Statistical Methods of Phylogenetic Analysis:

Including Hadamard Conjugations, LogDet Transforms,

and Maximum Likelihood

A thesis presented in partial fulfillment
of the requirements for
the degree of Ph.D. in
Biology at Massey University

Peter J. Waddell

1995

Abstract

This thesis studies phylogenetics from a biological-statistical perspective. Chapter 1 offers an overview of the field, with particular emphasis upon the classification and interrelationships of phylogenetic methods. Separating tree selection criteria from 'corrections' for multiple hits is crucial to understanding the behaviour of different methods. Chapter 2 extends Hadamard conjugations to allow for a distribution of unequal rates at different sites in a DNA sequence. This can be done, with minimal additional computational effort, assuming a gamma, lognormal etc. distribution of site rates. The result is either 'correction' of observed sequences assuming a certain distribution of rates, or prediction of sequence probabilities given a distribution of rates and a tree. A new set of faster Hadamard conjugations for correcting four state data are presented. These conjugations also allow unequal rates across sites, transition to transversion weighting and fixing the transition to transversion ratio.

Chapter 3 considers the more general time reversible and LogDet-Paralinear distances. These are extended to accommodate unequal rates across sites. It is shown that removing a proportion of constant sites gives the LogDet a high degree of robustness to unequal rates across sites even if the true model is not invariant sites plus identical rates. Analyses of 16S-like rRNA with constant site removal (CSR) LogDet reveals surprising results, including good evidence that Microsporidia are the most distantly related (i.e. first branch) eukaryotes. Chapter 4 deals with understanding the sampling properties of transformations, especially the Hadamard conjugation. Results include forcing the Hadamard conjugation to the Kimura 2ST and Jukes Cantor models, thereby reducing sampling variance. In doing this families of tree informative linear invariants were found. It is also shown that replacing log functions with truncated power series can reduce sampling errors (RMSE) substantially.

Chapter 5 deals with tree selection criteria. Studies reveal some interesting inter-relationships between Hadamard conjugation, distance and maximum likelihood (ML) based methods. Calculation of likelihoods with unequal rates across sites (e.g. a gamma distribution) are also developed. This can be done quickly with Hadamard conjugations, and a variety of sequences and models are studied. ML solutions to inferring reticulate phylogenies are described, and in an application are used to infer the population size of our ancestors with chimps and gorillas. A wide variety of methods, including ML, are shown to be inconsistent in the Felsenstein zone when site rates are unequal (in a similar situation ML is also seen to be inconsistent under a molecular clock). Overcorrecting the data is also a potential pitfall, and the concept of the 'anti-Felsenstein zone' is introduced, illustrated, and developed. A related phenomena is that two or more optimal binary trees can predict exactly the same sequences when rates across sites are unequal, and examples are provided. Chapter 6 describes new statistical tests. These include faster model based resampling to evaluate fit of model to data and tests of whether two data sets came from the same tree. A Bayesian view of support for different trees is presented. The thesis is large, but well illustrated, and looking at the figures alone should provide a useful overview of new results.

Acknowledgments

I would firstly like to thank my supervisors Professors David Penny and Mike Hendy for inviting me to participate in the phylogenetic research at Massey University, and allowing me to pursue a wide field of study. Both have been generous with their time and patience over a long period. David's enthusiasm for biology is always encouraging, as is Mike's love of mathematics. Other members of the group, especially Peter Lockhart and Mike Steel have provided assistance and encouragement in too many ways to mention. Thanks to you all.

Many statisticians have indulged my interest in the topic, especially Greg Arnold (a co-supervisor) and Terry Moore, and at other important times Brian McArdle and Chris Triggs. Graeme Wake helped tremendously by hosting me for much of the time in the Department of Mathematics and Statistics. Thanks also to the people I meet on my sojourn to the USA, especially David Swofford, Jaxk Reeves, David Hillis, Dick Hudson, Joe Felsenstein, Terry Speed, Walter Fitch, Jeff Thorne, Nick Goldman, Mike Donoghue, Arend Sidow, Mitch Sogin, Masatoshi Nei, Ken Kidd, John Hartigan, Adrian Gibbs and Rebecca Cann,. Your hospitality and enthusiasm are well marked.

Special thanks to Trish McLenachan, without whose encouragement and proof reading this thesis might never have been finished.

My family were always supportive in the background, which unfortunately is the way it has been for a long time now; certainly one of the personal costs of the work.

To everyone else who has aided me through this long period I extend my thanks and apologize that I cannot mention you all here. I look forward to rejoining society and enjoying your company once again.

Contents

4 Sampling errors associated with transformed data

Glossary of symbols and abbreviations

Generally, special symbols have a common meaning throughout the thesis, although some, of necessity, have multiple uses. If a generally defined variable such as *c* (usually being the sequence length) is used in a different context (e.g. as the shape parameter of the Weibull distribution), then this second usage will be indicated specifically in the text. A short list of symbols specific to chapter 3 is given at the end of section 3.1.

Symbol	*Definition*
α	The probability of a type 1 error in a statistical test (that is, rejecting the null hypothesis when it is correct).
δ	A transformed distance, or an additive distance (if in bold a distance matrix) (except appendix 3.1, where it refers to the delta method approximation)
γ	A vector description of pattern frequencies taking account of the effect of multiple hits with the Hadamard conjugation (or in the case of γ_D an approximation to this)
$\gamma(T)$	A vector description of a weighted tree
γ_D	A spectrum estimated from just pairwise distances
$\hat{\gamma}$	An estimate of γ based on a sample of sites
$\hat{\gamma}(T)$	A tree inferred from $\hat{\gamma}$
Γ	The gamma probability distribution (usually of the λ_j)
λ_j	The relative substitution rate of sites in set *j*
c	The sequence length (or shape parameter of the Weibull distribution, section 2.3.3)
c.v.	Coefficient of variation (standard deviation / mean)
d	The shape parameter of the inverse Gaussian distribution
d	A distance
d_{obs}	An observed distance
d.f.	Degrees of freedom in the χ^2 distribution
f	A vector of observed site pattern frequencies = *c***s**
k	The shape parameter of the Γ distribution
kb	One thousand nucleotide base pairs
G^2	The log-likelihood ratio goodness of fit statistic (the G statistic of Sokal and Rohlf, 1981).
H	A Hadamard matrix
i.r.	Identical rates of substitution across sites
M	A moment generating function, e.g. $M_\lambda(t) = E[e^{\lambda t}]$, with inverse M^{-1}
ML	Maximum likelihood

p_{inv}	A proportion of invariant sites (sites which cannot undergo substitution)
$\mathbf{s}$	A vector of observed site pattern probabilities (which sum to 1) = $\mathbf{f} / c$
$\mathbf{s}(T)$	A vector of observed site pattern probabilities generated by a particular tree evolutionary model
$\hat{\mathbf{s}}$	Observed site pattern probabilities (proportions) estimated from a sample
s.d.	Standard deviation
t	The number of taxa (or a dummy variable for a moment generating function in chapter 2, or a time scalar in chapter 3, as specifically indicated)(as a superscript to a matrix it means transpose)
tr/tv	Transition to transversion ratio
T_{12}	The tree ((1,2), 3, 4)
T_{star}	The unresolved tree (1, 2, 3, 4)
$\mathbf{V}$	A variance-covariance matrix
var	Variance

equifrequency	The states are in equal proportions
OLS	Ordinary (unweighted) least squares
WLS	Weighted (usually by the inverse of the variance) least squares
GLS	Weighted least squares, taking account of correlations
SS	Sum of squares
URAS	Unequal rates across sites

CHAPTER 1:

ESTIMATING EVOLUTIONARY TREES FROM DNA SEQUENCES.

"Nothing in phylogeny makes sense except in the light of the analysis."

1.1 INTRODUCTION

Phylogenetic trees are to evolutionary biologists what maps are to geologists, yet concerted study of their estimation is a surprisingly recent development. Perhaps the greatest boost to the prominence of phylogenetic methodology has been the recent advent of abundant sequence information. Since 1965 it has been recognised that polypeptides contain a history of past alterations which may be used to estimate the relatedness of different sequences (Zuckerkandl and Pauling 1965). Advances in sequencing technology and methods of sample preparation (e.g. the "Polymerase Chain Reaction" or "PCR") have put sequence data at the disposal of many biologists, and resulted in vast numbers of sequences available from data bases such as Genbank and EMBL. This thesis deals exclusively with methods of analysing such sequences with the aim of recovering information about their history, that is their evolution. By analogy some results reported here are also relevant also to similar studies of other data sets, e.g. electrophoretic, biochemical or morphological data.

When lineages seldom recombine, but rather evolve apart, the natural way to depict their relationships is by a tree of relationships. Over the past two decades many algorithms have been developed to estimate trees from sequences (see Felsenstein 1982, 1988, Swofford and Olsen 1990 for reviews). Unfortunately, due partly to computational complexity, few have been based upon statistical criteria. Further, even these have often been compromised by the sparse nature of DNA sequence patterns, making asymptotic tests both uncertain, and sometimes misleading (by sparse we mean many patterns with a low frequency of occurrence). In this thesis we will develop a statistical framework centered around Hadamard conjugations, a new method of correcting sequences for multiple changes prior to, and independently of, tree selection. A real benefit of working with Hadamard conjugations is that they often shed light on the statistical nature of other, frequently used, methods of inferring trees. It is desirable to gain a deeper theoretical understanding of these methods, as many assessments to date have been based upon simulations that lack generality.

1.2 THE RATIONALE FOR STATISTICS

1.2.1 Popperian ideals

One philosophy has come to stand out in the latter 20th century as an ideal by which we gain knowledge, as opposed to just a description of observations. This ideal, often associated with Karl Popper (thus called "Popperianism"), is a prescription for good science, or any other discipline interested in uncovering some aspect of reality (e.g. see Popper 1979). It basically states that we do not have absolute knowledge, but the best way to move forward towards a better understanding of the world is to:

(a) Construct hypotheses, which are specific enough in space and time, that we expect to be able to gather data which will disprove at least most of them.

(b) Collect the most relevant data possible and make tests of the previous hypotheses, with a view to falsifying all those that do not pass the test.

(c) Review the remaining hypotheses and seek modifications of any which are logically internally consistent, and are not falsified by the presently available data. These revised and sometimes quite novel hypotheses are then subjected to (b) and so on (Popper 1979). In this way we expect to move towards reality (we use 'expect' to mean overall, rather than necessarily at every step) and avoid the pitfall of more and more constructing our own "reality", which we become fools to.

Because this is a prescription of how science is often done, it is quite consistent with the sociological science descriptions of Kuhn and others (e.g. Kuhn 1977). Popperians are particularly interested in speeding up the processes by which we gain knowledge (Popper 1979). An important step is the generation and selection of new hypotheses, which in a sociological context must include how we educate the (young) people that will supplement the growing body of knowledge. Another important step that is attracting much deserved attention is the process of testing. As we mentioned earlier we do not have absolute knowledge, and as such our falsifications are not absolute either. In such a case it is natural to look to the concept of probability and the associated area of statistics for guidelines. An especially useful aspect of statistics is its ability to attach probabilities to chance outcomes, based upon a statistical model. If our model is reasonable (which must itself be tested) then when we have outcomes that are unlikely to be due to chance they constitute good evidence with which to test alternative hypotheses. Conversely if there is no specific hypothesis, unlikely outcomes provide an ideal starting point for a new hypothesis (a common situation in a new field such as molecular biology).

1.2.2 The question that Bayes tried to answer

One of the most elusive areas in the philosophy of science has been deciding on how to assess quantitative data so as to come to a qualitative decision, namely the rejection of a hypothesis. Quantitative decisions are important in that they bring a degree of coherency and precision to science not otherwise attainable. That is not to say that science doesn't work without

them, but the aim is to make science more efficient in its testing of hypotheses, hopefully getting answers quicker and with more consensus. Apart from the benefit of forcing us to be specific in the construction of our hypotheses, such tests offer a way of dealing with the random component in all answers, a way to assess data which never exactly agrees with expectations.

Bayesian theory offers insights into part of the process, but it is not without controversy, as some statisticians still distance themselves from colleagues that take this approach. And even in the philosophical-statistical area of decision theory much disagreement still exists (e.g. Maher 1993). One approach that continues to look promising is the modification of a prior probability with an experimental likelihood (the probability of the experimental outcome given hypothesis a), which is in essence Bayes theorem. Bayes theorem states that the posterior probability varies as the prior probability multiplied by the likelihood, or

Posterior probability(x) = Prior probability(x) × likelihood(x).

Bayesian approaches while able to be made theoretically coherent, have at least two major problems (calculations aside) when applied to the most interesting scientific questions: 1) We don't know which model to evaluate the likelihood under and different models may give substantially different answers, and 2) Most scientists cannot agree on a prior probability (Smith 1993). The first problem is of course common to all statistical procedures, arguably less so to the Bayesian approach than to the standard "frequentist" approach to hypothesis testing, since possible error in the experimental evaluation is diluted by the prior knowledge. In either case, within a field modeling, selection is usually addressed by a progressive extension of mathematical models, combined with prudent ways of selecting the best performing models or parts of (some would call these the most parsimonious models, that is those that best explain the data with the fewest parameters). In addition we also have substantial input from related areas of biology such as the mutation chemistry of DNA (although often the greatest confounding factor here is that assessment of how a particular protein has evolved requires a reliable estimate of its evolutionary history). Hopefully we end up with a set of "reasonable models" which substantially agree in their evaluations of likelihood for a given set of data. Phylogenetics appears to be making good progress, with extensions to models appearing ever more frequently (e.g. Cavender 1978, Felsenstein 1981a, Hasegawa *et al.* 1985, Barry and Hartigan 1987a, Kishino *et al*. 1990, Adachi and Hasegawa 1992, Yang 1993, Felsenstein 1993, research in this thesis).

The second problem of getting people to agree on prior probabilities is more difficult, and is sometimes the result of individual idiosyncrasies. It is however true that individuals often agree much better when they have to make decisions on priors which carry penalties if the probability is either too small or two large (for some useful approaches see Stuart and Ord 1987, p.263). As long as we are choosing between a small number of trees (e.g. three trees for four taxa, or fifteen for five taxa), then the conservative approach of giving them all the same prior often does not overly hinder coming to conclusions. As yet Bayesian type approaches have found little favour in phylogenetics, yet it is just this sort of objectified statistic that is needed to be able to make an objective appraisal of the reliability of phylogenetic claims. Informativeness to non-phylogeneticists should be enshrined as a primary aim of phylogenetics; the ultimate rational for

reconstructing trees is not to perpetuate the process of reconstructing trees, but to supply all biologists with essential information about biological history. Further, to quote Stuart and Ord (1987, p284): "The problem that Bayes attempted to solve is supremely important in scientific inference and it scarcely seems possible to have any scientific thought at all without some solution of it, however intuitive and however empirical." All the better then if our solution is generally well behaved and widely agreed upon.

Some authors (e.g. Edwards 1972, 1992) have tended to reject the prior probability as not being validly obtainable, and rather attempted to develop a coherent framework around the experimental results by themselves. We will consider Edwards' favoured solution of careful exploration of the likelihood surface further in chapter 6.

In this thesis we will argue that a form of Bayesian inference is highly desirable in a rapidly expanding field such as molecular biology, as it should give an up-to-date estimate of how decisive the relevant data is. It would seem desirable to replace the present, often unjustifiable urge to be the "person to prove a hypothesis" with a more realistic window on reality, the intermittently declining probability of all present hypotheses as they evolve, with the more useful hypotheses leaving descendants as science progresses. That is a sort of running tally of where we are at this point in time, combined with recognition of the data and theory that have got us there. Bayesian estimation serves as a counterpoint to the intellectual equivalent of the "knockout punch", the all or nothing hypothesis test (all the more problematical as it is often based on a model with assumptions we know are violated).

1.3 TERMINOLOGY ASSOCIATED WITH EVOLUTIONARY TREES

The process of sequence evolution modeled here is sequential separation into distinct lineages, which do not rejoin. Such a process is described exactly by a diagram or graph (from mathematical graph theory) called a *tree*. Penny *et al.* (1992) argue that different branches of science should use the same notation where ever possible to avoid multiplicity, and phylogenetics should not be an exception. For example a "*cladogram*" is a tree, yet this alternative term became popular with the Hennigian view of phylogenetics in the 1970's and 80's. Wherever possible terms should be descriptive, that is have a commonly understood usage. With this in mind a coherent set of labels is now presented for use in this thesis, although we list commonly used alternatives.

Figure 1.1a shows a tree associated with the notation used within this thesis. The tree is made up of *nodes* (points or vertices) linked together by *edges* (internodes, links, lines or arcs). The *degree* of a node is the number of edges that run into it; a term we will rarely use in this thesis. A tree is *binary* if all internal nodes separate a pair of lineages at a time, implying that all internal nodes (excepting possibly the root) are of degree 3. A *star* tree, in contrast, has no internal edges, with all external edges coming from one internal node. A non-binary tree that is not a star tree, is termed a *partially resolved tree*, on the biological expectation that sequences separate in pairs, since the probability of two mutations, becoming fixed (see below) from the same parent is

extremely low. The *tips* (terminal nodes, leaves, or pendants) of a tree (here drawn as rooted and upside down, implying time flows forward reading down the page) are associated with a subset of the data, often a sequence from one individual. If the time separating a sequence from its nearest relative in the tree is large in comparison to the effective (long term) population size of that sequence, then the tree is unlikely to be different when we choose other sequences from that population. Accordingly, the sequences may also be labeled *taxa* (sometimes also operational taxonomic units, OTU's,) when associated closely with particular organisms of a distinct nature (e.g. species).

One particularly useful combination of terms is *external* or *pendant* (as opposed to internal) to denote an edge leading to just one tip. Such edges do not enter into describing the unweighted tree, but their weights (or lengths) are biologically important. In phylogenetics, one especially variable term is *branch.* In botany, and frequently in mathematics, it refers to the subtree obtained by cutting a tree along an edge. For rooted trees this is the subtree not including the root, for unrooted trees convention usually indicates it is the smaller subtree. On an *unrooted tree* (see figure 1.1b) such a term may be made exact by specifying the edge at which the cut is made and naming just one tip of the branch. The usage here is both botanically and mathematically correct (in agreement with Penny et al 1992), as opposed to using this term to describe an edge or sometimes a path through a tree.

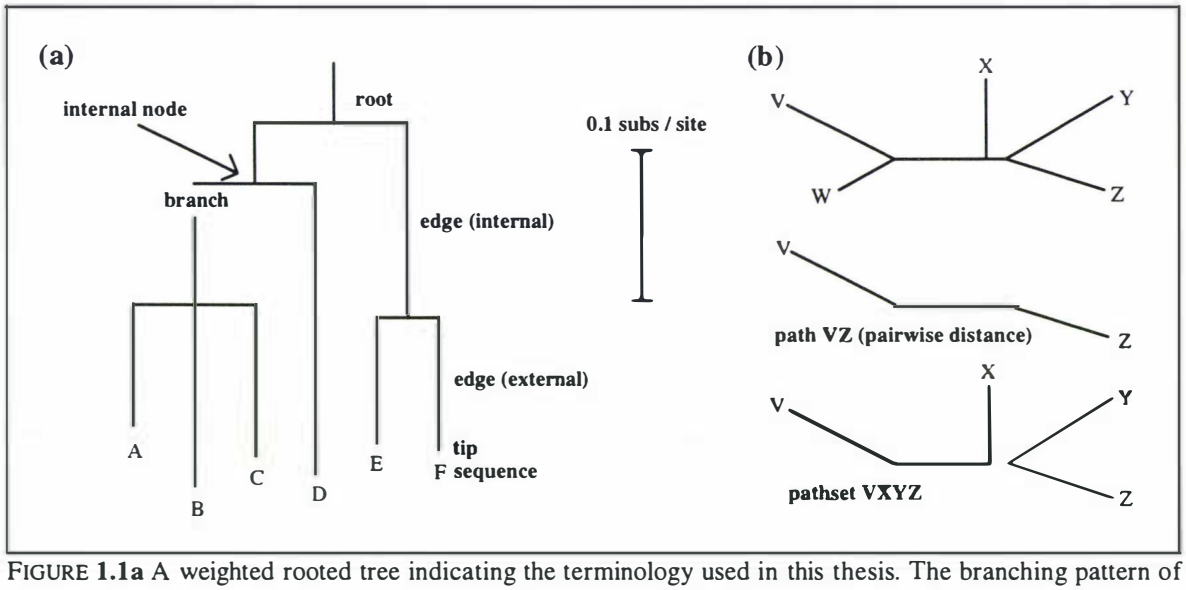


FIGURE **1.1a** A weighted rooted tree indicating the terminology used in this thesis. The branching pattern of group (A, B, C) is unresolved. This tree can be written in parenthetical notations as (((A, B, C), D), (E, F)). Notice that the horizontal lines are just conveniently sized spacers, all information on edge lengths is conveyed by the vertical lines, which are calibrated by the scale. **1.1b** An unrooted but weighted tree illustrating what we mean by a path and a pathset (as used in a Hadamard conjugation).

A useful classification of trees is into the forms *rooted unweighted, rooted weighted, unrooted weighted,* and *unrooted unweighted.* Most tree building methods simply construct an unrooted tree. Typically, sequence trees are rooted by including an "*outgroup*" sequence(s) from an organism(s) "known" to fall outside of the group of sequences being rooted (i.e. none of the outgroups have a more recent common ancestor with any *ingroup* taxon, than the ingroup taxa share with one another). For accuracy it is desirable to have more than one *outgroup* sequence.

Hopefully the sequences will be close to the group being rooted, but not so close as to mislead the analysis due to ancestral polymorphism (i.e. the sequence tree does not match the species tree due to allelic diversity).

Edge *weights* (or *lengths*) are most often measured in terms of how many nucleotide *substitutions* occurred on that edge. Most commonly it is the *expected* (average) *number of substitutions per site* and this measure takes into account multiplicity of changes at some sites. In this thesis it is explicitly stated whenever we are referring to the *observed* number of substitutions per site, which by definition will hide any multiple changes at a site. The observed number of changes, implied by mapping the sequence data onto the true tree, is always less than or equal to the actual number of substitutions. The actual number of substitutions is additive on the tree, but the observed number is not tree additive as soon as any site has multiple substitutions (by additive we mean that all the pairwise distances measured in this way can fit exactly on the same weighted tree). Both measures meet the less stringent criterion of being metric (that is $d_{ij} \geq 0$, $d_{ij} = d_{ji}$, and $d_{ij} \leq d_{ik} + d_{kj}$ for all i, j, and k, where d_{ij} is the observed number of substitutions between two points, and k is any other point on the tree; the so called triangle inequality property). Traditionally defined, a substitution occurs when an DNA replication error (a *mutation*) becomes ubiquitous (or *fixed*) in a population, so that it appears in all descendants of a lineage. Generally we will be dealing with clearly distinct populations, so we need not be concerned about defining substitutions within populations.

This thesis often deals with four taxon unrooted trees, so a labeling such as T_{13} means an unrooted tree with taxa 1 and 3 closest together (which also implies 2 and 4 are together). If a four taxon tree is unresolved, that is has an internal edge of length zero, then it is called a *star tree*, T_{star}. For more than four taxa we use the standard bracketed notation to describe the hierarchical groupings of a tree (for example see the Phylip 3.5 Manual, Felsenstein 1993).

As used in this thesis, a *path* is the direct route between two tips. If it is the actual distance (counting multiple changes) then its weight (*path length* or *pairwise distance*) is the sum of edge weights along the path. From a complete matrix of such exact *tree distances* we may reconstruct the tree unambiguously using consistent algorithms (see below). As with the observed number of changes, *observed distances* are never greater than the tree or true distances, the relationship between the two being a *monotonic non-linear* transformation (at least for Markov models with independence between sites, the only models presently studied in depth). A matrix of tree distances is *additive*, that is, they will fit a weighted tree exactly. Unfortunately additivity of distances does not guarantee consistency, as in some circumstances an apparently valid transformation can make observed distances additive, but upon a different tree to the one that generated them (an example of this is in chapter 4)! Alternatively *non-additivity* may be due to sampling error and/or moderate bias, and need not prevent us from identifying the (unweighted) true tree with the data at hand, and a robust tree building algorithm.

Hadamard conjugations extend beyond just simple paths or pairwise distances. A *pathset* is the set of paths, with no edges in common, connecting pairs of taxa (e.g. see figure 1.1b). A pathset weight (or length) is the sum of the weights of the edges in that pathset. Notice then that

a pathset between four taxa (ijkl) is equal to the minimum of $d(T)_{ij} + d(T)_{kl}$, where d(T) is the true tree pairwise distance. Similarly for larger pathsets e.g. the pathset weight for a sextet (pathset with six endpoints) is the minimum sum of three distances, connecting pairs of taxa from the sextet. As such, the length of pathsets may be considered *generalised distances*. Under some models these pathset weights may be calculated from frequencies of tip patterns alone (e.g. as used in sequence Hadamard conjugations, chapter 2), without needing to identify the minimum of a sum of pairwise distances.

1.3.1 A first taste of Hadamard conjugations

The essence of the Hadamard conjugation is that for certain mechanisms of nucleotide substitution there are as many directly estimable generalised distances (pathset lengths) as there are distinct sequence site patterns. A sequence pattern is the pattern one sees when looking down a column of aligned sequence data e.g. with four taxa, a pattern might be RRYY (where R is a purine base, and Y a pyrimidine) or in greater resolution AGCC (where the letters stand for the nucleotide bases). By using an invertible linear transform between a vector of sequence patterns and the generalised distances, we can change the one into the other, make corrections for multiple substitutions based on long sequence expectations, and then reconstruct the expected magnitude of all sequence site patterns once the effect of multiple substitutions is accounted for. Tree selection algorithms such as compatibility, closest tree, or parsimony can then be applied to such corrected sequence data, with the potentially misleading effect of multiple substitutions negated (under the model, with long enough sequences).

Here we look at the context of Hadamard conjugations as they relate to other methods of sequence analysis, while the mechanics and notation of this class of methods are presented in chapter 2. Because Hadamard conjugations represent a bridge between distance methods and sequence methods of phylogenetic analysis, they are cast in a variety of different contexts in this thesis. Hadamard conjugations are also intimately tied up with the concept of "invariants", and are discussed in this context in the section on current methods of phylogenetic analysis (later in this chapter). They are also used to shed light upon the nature of current maximum likelihood methods used in analysing DNA sequences (this is shown particularly in chapter 5). For the moment all the reader need note is that they correct sequences for multiple hits, or can be reversed to predict all sequence pattern probabilities given a weighted tree.

This thesis also uses a moderate amount of mathematical notation and here we define conventions adhered to throughout. Our notation for a matrix is a capital in bold, e.g. $\mathbf{H}$ (with individual entries in plain text e.g. H_{ij}), and for a vector lowercase in bold, e.g. $\mathbf{s}$ (with entries s_i). Due to an important precedence two vectors have Greek symbols, namely $\boldsymbol{\gamma}$ and $\boldsymbol{\rho}$. Unfortunately the bold symbol font available for the thesis does not stand out, so wherever the symbols γ and ρ appear in the text without subscript, they should be taken to refer to the complete vector. We use the abbreviation *ln* to indicate the natural logarithm function (log to base e), and it is applied in turn to each individual component in its argument (that is componentwise) unless otherwise stated.

1.4 STOCHASTIC MODELS OF EVOLUTION

A tree model (hereafter just a model) of sequence evolution has four main parts:

(1) A tree (rooted or unrooted depending upon the mechanism of change).

(2) A set of edge weights.

(3) The initial frequencies of the character states at the root of the tree.

(4) A mechanism of character evolution. This includes a transition matrix determining the probability that nucleotide i at site x changes to nucleotide j along a given (weighted) edge of the tree.

Parts one and two can be combined into one (i.e. a weighted tree) but it is often conceptually useful to keep them separate. For example, is a tree selection method good at recovering the unweighted tree under a certain model? And if so how well does it do at estimating edge weights? Another way of thinking of tree inference is that once we have settled upon the concept of a tree (as opposed to a network, a chain of being or other possible descriptions, e.g. see Penny *et al.* 1994) then we are looking for the tree model which best describes the data. In making such evaluations it is important to consider how (4) the assumed mechanism of change alters our perception of the "best" tree.

This thesis deals predominantly with "independent and identically distributed" (i.i.d.) Markov mechanisms of character change (see for example Cavender 1978, Felsenstein 1981a, Keilson 1979). A general term for a stochastic model is a Markov model (Keilson 1979). One way of classifying Markov models is by how many previous states need be considered to estimate the probability of the next change. If we assume independence of sites, then under our models of evolution, the current state tells us everything about the probability of the next change. This is accordingly called a first order Markov model (the order being how many previous states need be known to evaluate the probability of the next change). If the probability of the next change can be estimated without even knowing the current state (for example all substitutions are equally likely), then this is called a zero order Markov process.

An i.i.d. evolutionary model is one in which evolution at all sites is independent and identically distributed (e.g. Cavender 1978), that is sites will undergo substitution independently of one another, but all sites follow a common underlying process of substitution. One generalisation of earlier evolutionary models is to drop the identical rates (i.r.) at all sites assumption and instead give each site an intrinsic relative rate; that is the probabilities of site y changing is a constant multiple of the rate matrix for site x. We do not need to estimate the rate at every site, but can specify for example, that the true distribution of rates across sites follows a standardised and normalised gamma distribution, with just one shape parameter as the free variable to optimise. Chapter 2 looks at a variety of distributions of rates across sites and shows how such mechanisms of change can be married to Hadamard conjugations (while in chapter 5 we use these findings to perform maximum likelihood calculations). Such models tend to fit the sequence patterns generated during the evolution of coding regions much better than identical

rate models, due to unequal evolutionary constraints at different sites. Because sites are still independent and if you consider that the rate of any site is randomly sampled from the same underlying distribution of rates across sites, then it is often considered that this is still an i.i.d. model. In this thesis we will refer to these models as non-i.r. (non identical rate) but still i.i.d.

It is important to limit the number of parameters needed to describe the rates at different sites. Models which do not constrain the number of free parameters may become inconsistent, especially if the number of parameters grows linearly with the amount of data (Stuart and Ord 1990, p. 682). Felsenstein (1973) for example, argues that this was the problem with Farris' (1973) maximum likelihood justification of parsimony. However if we can place restrictions on the distribution of rates across sites on each edge of the tree (for example, that sites are fixed in their rates relative to one another), then we could have an accurate model with only a few extra parameters. For such an i.i.d. model we could evaluate the likelihood of a sequence pattern, given a weighted tree and a description of the process of substitution, as a sum of the probability of this pattern over all possible internal node assignments and over all allowable edge lengths (or an integral for a continuous distribution of site rates)(Felsenstein 1981a). Such a model may be consistent as the number of parameters can easily be less than the number of distinct sequence patterns. Such models are developed and evaluated in this thesis.

Interestingly Penny *et al.* (1994) have shown that for 2-state characters parsimony methods are maximum likelihood estimators under conditions where we cannot put any constraints on the relative rates of different sites. One such model is where the rate of substitution at each site changes independently of all others, and that we know nothing about what its new rate will be, even though the transition process itself may remain identical (e.g. Poisson, also known as the Jukes-Cantor model with 4 character states). This model implies that sites change their rates relative to each other, so a site could have a unique rate on every edge in the tree. Under these conditions parsimony, which in this instance is also maximum likelihood, can be inconsistent (Penny *et al.*, 1994). Fortunately with biological sequences, the general homogeneity of the enzyme repair mechanisms for sites along a sequence will usually preclude the need to consider such general models in tree estimation (although they may become more prominent when site substitution is governed by selective forces).

Important terms used with Markov tree models are *homogeneous* and *nonhomogeneous*. A homogeneous model is one in which the relative probabilities of all substitutions are the same on all edges of the tree. For example, if there are just two rates defining all substitutions, say transversions are at rate 0.1 and transitions at rate 0.4, then tr / tv = 4. Another edge with transition rates 0.2 and 0.8 is homogeneous with respect to the first edge, but an edge with rates 0.1, 0.5 is not. Each transition matrix will drive the base composition (or states) of a sequence towards a single optimum, which is dependent upon the form and numbers in the transition matrix of substitution probabilities. With a *nonstationary* model, the frequencies of states (i.e. base composition) will vary through time. There are two possible reasons for this: The initial frequencies of states (i.e. at the tree's root) where not at equilibrium with the substitution process on edges of the tree, or the model is nonhomogeneous, with different transition matrices

implying different stationary base compositions. Under a *time reversible* stochastic i.i.d. model of evolution the root of the tree can be placed anywhere, and the probability of the data (the site pattern probabilities) will remain the same. Two things which usually (but not always) make a model *non-time reversible* are a root base composition not in equilibrium, or a nonhomogeneous substitution process. A third possibility for a non-time reversible process is a stationary, homogeneous model, where the transition matrices are of a non-time reversible form (this is described in more detail in section 3.2).

In this thesis we will be primarily considering a subset of symmetric rate matrices (the product of an instantaneous rate matrix and a scalar, often called *time* but best thought of as an amount of evolution not necessarily linear with time). The symmetry of rate matrices guarantees that as long as the root distribution is at the substitution processes equilibrium value, then the probability transition matrices are also symmetric. This in turn implies that evolution is time reversible, therefore the substitution process itself cannot be used to root the tree; Felsenstein (1981) discusses this point. The most general substitution scheme that the 4-state Hadamard conjugation can model is labeled the generalised Kimura 3ST model (Kimura 1981, generalization by Evans and Speed 1993). Each edge in the tree has a rate matrix of the form,

$$\begin{array}{c} \\ A \\ C \\ G \\ T \end{array} \begin{array}{c} \begin{array}{cccc} A & C & G & T \end{array} \\ \begin{bmatrix} - & \psi & \alpha & \beta \\ \psi & - & \beta & \alpha \\ \alpha & \beta & - & \psi \\ \beta & \alpha & \psi & - \end{bmatrix} \end{array}$$

where α is the transition rate, β the transversion type 1 rate, and ψ transversion type 2 rate. This model may be conditionally non-homogeneous, that is the values of its parameters can vary freely from edge to edge, but the matrix retains its specified structure. When $\beta = \psi$ on each edge in the tree we have the generalised Kimura 2ST model (based on Kimura 1980), while if $\alpha = \beta = \psi$ on all edges of the tree we have the Jukes-Cantor model (Jukes and Cantor, 1969). If all edges of the tree have α, β and ψ in a fixed ratio, we call this the standard Kimura 3ST model (Kimura 1981), while $\alpha = \beta$ and ψ in a fixed ratio give the standard Kimura 2ST model (Kimura 1980). This model (and its submodels) have the unusual property that they can be mapped down to 4^{t-1} pattern probabilities which will remain identical no matter where the root is placed, or what base composition it is given (so in a sense they can define a *time-reversible* model under this circumstance, although when using all 4^t data patterns they behave like all other models).

This thesis also considers more general i.i.d. models. If the instantaneous rates of substitution are not identical across the tree, then we may need to model up to 12 parameters per edge in the tree even if sites remain i.i.d. and i.r. Such a model was implemented by Barry and Hartigan (1987a). Notice that there are at most (2t - 2) times 12 parameters in this model (when the tree is binary and rooted), which is always less than the number of site patterns possible (4^t), as long as $t > 2$. This point being in agreement with Blaisdell's (1985) observations on recovering substitution rates from a pair of sequences given only their pairwise divergence matrix. This is also consistent with the findings of Steel (1994a), who showed via the properties of the determinants of Markov transition matrices, that every possible i.i.d. and i.r. model produces a unique sequence spectrum (so with long enough sequences correct tree reconstruction is always

possible under this model). Chapter 3 of this thesis considers in detail a distance method that will be consistent under any i.i.d. and i.r. model, this being the LogDet / Paralinear transformations of Steel (1994), Lockhart *et al.* (1994), and Lake (1994). This distance is very general, and applies to all i.r. and i.i.d. nonhomogeneous, nonstationary, and non-time reversible models. This work includes: New understandings of how to interpret this distance. Making this transform more additive (and hence robust) when there are unequal rates across sites. Comparing its performance in combination with tree building algorithms, with other tree estimation methods when analysing a 28 taxon data set pertaining to "the tree of life." This chapter also extends general time reversible distances, which are exact under stationary and homogeneous models with up to nine free parameters, to allow for a distribution of rates across sites.

The last point made here is the nomenclature of a variable used in some Markov models, and also generally throughout this thesis. To call a site invariant in this thesis implies that it cannot change, while unvaried simply says a site is constant and shows no evidence of having changed although it may have the potential to do so. Usually these two terms should be kept separate. When using current methods to estimate of the proportion of invariant sites, it is uncertain if this is what is really being estimated, rather than just the optimal number of unvaried sites to treat as invariant in order to optimise some fit of data to model. Thus this thesis tends to use the term p_{inv} to simply indicate the proportion (measured relative to all sites) of unvaried or constant sites removed (or otherwise treated as invariant) in a specific instance. A further distinction is that the term $\hat{p}_{inv}$ is used to indicate a specific, somehow optimal, estimate of p_{inv} made by a method.

1.5 STEPS IN RECONSTRUCTING TREES FROM SEQUENCES

In this section we look firstly at the overall structure of phylogenetic analysis of sequences, then give a list of desirable properties for tree reconstruction algorithms, and finally briefly review present methods of tree reconstruction.

1.5.1 The logical structure of phylogenetic analysis

Going from sequences to conclusions about the evolution of molecular sequences involves five main steps:

(1) Alignment of sequences, so that each column represents a homologous site.

(2) A method to deal with potentially misleading multiple changes.

(3) A tree evaluation criterion.

(4) A method of searching out an optimal tree.

(5) A means of evaluating the results statistically.

(1) Alignment: This thesis does not directly address the issue of sequence alignment. However this step is of critical importance in many studies, especially when using anciently diverged molecules with many insertions and / or deletions. In one of our data sets (specifically 16S-like rRNA) there has been controversy over different alignment methods, specifically those

of Lake (1987), verses the more usual procedures of Dams *et al.* (1988), Olsen and Woese (1989), and Gouy and Li (1989a). In one of the analyses in this thesis we can see the effects of these different alignments on a Hadamard conjugation, which cautions us regarding the bias that alignment errors can cause in the final analysis. An important issue which is still in its infancy, and which we do not pursue here, is the matter of iterative sequence alignment involving tree estimation, then sequence alignment using the tree, and so on (e.g. Hein 1990).

(2) Negating multiple hits: Step (2) addresses the question of how to deal with multiple substitutions in the data. The advantages of making no transformation of the original data are computational simplicity, and no increase in the variance of the data. In contrast non-linear transformations of the original data invariably exaggerate sampling errors. Consequently tree selection will sometimes do better on uncorrected distances (e.g. Saitou and Nei 1987, Hillis *et al.* 1994) and sequence patterns (e.g. Charleston 1994, and chapter 4 this thesis). Not making a transformation results in systematic errors in estimated distances, perhaps not enough to mislead the tree selection criterion completely, but certainly enough to underestimate edge weights. An important decision to make at this point is whether to reduce sequence data to just the pairwise distances in order to use a large set of model exact distance transformations. Once the data are reduced to pairwise distances they cannot be converted back to sequence patterns, and this raises the question of how do distance based methods perform relative to the different types of sequence based methods (e.g. Penny 1982)? We consider this question especially in chapter 5.

The term *corrections* is used in this thesis to mean any modification with the aim of compensating for multiple hits. This does not necessarily infer that such corrections will be perfect with either finite or infinite data. Different methods make these corrections in different ways. Distances, invariants and Hadamard conjugations generally make tree independent transformations of the data (or a condensed form of it). A tree independent transformation aims to make the model parameters (tree edge weights for example) additive with respect to the data, so they are like the data transformation methods in general statistics (e.g. Box and Cox 1964). Tree selection following a transformation to restore additivity is a special case of a generalised linear model. In contrast the method of maximum likelihood works in the opposite direction, by starting with a fully defined model, then explicitly calculating the effect of multiple changes, so as to predict what the data would be like if that model were true. Thus a maximum likelihood model makes the effects of the model directly interpretable in terms of the probability of the data. In either case the aim is the same, to circumvent the misleading effect of multiple substitutions at a site. How well this is achieved by the different methods without perturbing other aspects of the analysis (e.g. the variance of estimated parameters) is something little understood in the context of tree estimation.

Chapter 2 of this thesis extends the Hadamard conjugation to corrections for mechanisms of evolution that allow rates across sites to vary. Maximum likelihood methods and distance methods generally are also model exact. Other methods such as weighted parsimony (e.g. Maddison *et al.* 1992), and more recently weighted likelihood (as implemented by Olsen 1994), aim to make modifications to improve robustness, although these corrections do not guarantee to

overcome the problem of parallel changes for a specified model with all possible parameter values. It is interesting to think of such ad hoc corrections as somewhat model independent, and certainly deserving of further study especially since they may sometimes do very well in tree reconstruction (e.g. Hillis and Bull 1993).

(3) Tree evaluation criteria: A variety of commonly used tree selection criteria are discussed below. Most of these criteria aim to maximise some measure of fit between the data and tree / model predictions, after calculations are made to cope with multiple changes (with no correction itself regarded as a choice). Likelihood maximises fit (for example by the likelihood ratio criterion) between observed data and that predicted by a tree for a particular mechanism of evolution. Distance methods usually minimise a measure of error (e.g. least squares) between the corrected distance data matrix and a tree. The *"least squares length"* criterion (Kidd and Sgaramella-Zonta 1971, Kidd and Cavalli-Sforza 1971), which was recently restudied, renamed and made popular as the "*minimum evolution*" criterion (Rzhetsky and Nei 1992a), minimises a slightly more indirect property-that of the sum of edge lengths when these are reconstructed by unweighted least squares. Parsimony and compatibility maximise certain types of fit (e.g. compatibility minimises sum of absolute deviations between data and tree, where deviations are often measured as the observed sequence pattern frequency). We look at different "classic" statistical fit criteria for use in tree selection from Hadamard conjugations in chapter 5, after describing the statistical framework of Hadamard conjugations in chapter 4.

(4) Searching amongst trees: Searching for the best tree is not insignificant, especially given that the number of unrooted binary trees grows as the double factorial (2t-5)!! (i.e. $1 \times 3 \times 5 \times ... \times (2t-5)$), which very quickly becomes a huge number (where t is the number of taxa, e.g. see Felsenstein 1978). Searching out the best tree by all the present "fit" tree selection criteria appears to be NP-hard (that is there is no known polynomial time algorithm guaranteed to work for all data sets, Graham and Foulds 1982). This area of phylogenetics involves numerical optimisation, and operations research problems. The approach of evaluating all trees explicitly soon falters due to the number of possible trees. About 10 taxa is usually the limit for any fit criteria, that is 2,027,025 binary trees, rising to 34,459,425 binary trees to evaluate with 11 taxa .

Branch and bound procedures (Hendy and Penny 1982) implicitly evaluate all trees by progressively excluding whole sets of trees with something in common which means that none of them can be optimal. Then explicitly evaluate what remains. This procedure is data dependent, but with good implementation has allowed the global optimum for up to 20 or so taxa to be found (Swofford 1993). Beyond this a whole spectrum of heuristic searches are possible, with few studies of which are best beyond recognition that more general rearrangements such as tree bisection and rearrangement (Swofford and Olsen 1990, Swofford 1993) reduce the chance of being trapped in local optima. Swofford and Olsen (1990) offer a useful summary of branch and bound and heuristic search procedures in phylogenetics.

Note that steps (3) and then (4) typically denote a double set of optimisation problems; finding the edge weights and sometimes other parameters that optimise the fit criterion for a

given tree (a numerical methods problem), then searching amongst trees for the best optimum (operations research problems). Methods which do not explicitly separate steps (3) and (4), for example neighbor joining (Saitou and Nei 1987, with the popular algorithm of Studier and Keppler 1988) or distance Wagner trees (see Swofford and Olsen 1990), can be more difficult to understand. Neighbor joining, for example, may be considered to evaluate trees approximately by the minimum evolution criterion at each step of an agglomerative (local) search procedure (Rzhetsky and Nei 1992a). A similar search procedure using parsimony would start with a star tree of all taxa, then taking each pair of sequences, evaluate which pair of sequences grouped together decreased the parsimony score by the most. This pair are then treated as inseparable, and the procedure repeated until only three groups remain. If two or more pairs give the same decrease in score on any one cycle, then the first pair only is grouped before beginning another cycle of agglomeration. This is sometimes called non-arbitrary tie breaking, and can lead to problems if the taxa are always labeled in a certain order. Without relabeling taxa, ties can also be broken randomly with a simple algorithm to keep track of the number of alternative arbitrary resolutions. Such a local search procedure may become less reliable as the number of taxa grows as the procedure does not search thoroughly among local optima and an error in a grouping early on could result in a rather different tree from what other search procedures might find. However studies such as those of Charleston (1994) on moderate numbers of taxa (20 to 40) show little sign of such problems with neighbor joining, suggesting that local search procedures can be surprisingly reliable (when the data fits the model reasonably well). Their other great asset is their speed (neighbor joining as implemented by Studier and Keppler 1988 is order t^3), which allows studies with many taxa, and / or analysis of many bootstrap samples for statistical purposes.

(5) Statistical testing. This is the penultimate step before drawing biological conclusions, and addresses questions of how certain the results are. The first question that one should ask is whether the assumptions of the analysis hold well enough that we can be confident the results are reliable indicators of relationships and not artifacts. Secondly, if they are not artifacts of systematic error, what is their probability of being due to sampling error? These last steps are often neglected, yet crucial to making a convincing presentation of results before a scientific audience. It is fair to say that until these last two questions have been answered the analysis is not complete, and this is why the development of statistical tests in phylogenetics deserves serious support. Chapter 5 looks at the foundation and extension of such tests.

By separating phylogenetic analysis into these five steps is possible to examine and modify the properties of each step in order to view the quality of the overall result. Take, for example, the method of parsimony. To some this conjures up tree selection upon the observed data and indeed this has been the setting in which it has traditionally been employed. However parsimony is in actuality a tree selection criterion (step 3), and as such an analysis utilising parsimony is dependent upon steps (1), (2), (4) and (5) as much as any other method. Looking at step (2), sometimes definite improvements can be made to site pattern data in order to suppress the effect of multiple hits (or, as the cladists call them, homoplasies). One approach is to use a direct weighting of the observed data to emphasis rarer changes, be they character state changes, or

changes in more slowly evolving characters. A more recent approach is the non-linear and model-exact reweighting performed by a Hadamard conjugation (Hendy and Penny 1993, Steel *et al.* 1993b, Charleston *et al.* 1994).

The choice of tree search criterion is also quite independent of the tree selection criterion, so with parsimony we can use anything from a very limited (and fast) local search, to searches which would take many years to complete. The extensive choices for tree searching available in the computer program PAUP (Swofford 1993) clearly illustrate that there are decisions to be made at this point. Algorithmic techniques such as neighbor joining often come in for criticism as not being reliable (e.g. Swofford and Olsen 1990). However they can do very well in simulations (e.g. Charleston 1994) and should not be dismissed as their speed can give them real advantages. It is important however that we study how neighbor joining works, seeking to understand how much of its good (or potentially bad) performance is due to tree search strategy verses tree selection criterion. If we knew this we would be able to make more objective and less subjective proclamations of its worth. This type of knowledge is urgently needed to make better predictions of how to tune steps (1) - (4) for particular types of analysis.

An important point which is still not fully resolved is how to ensure that step (5)(statistical evaluation) is adequately performed given the different choices at step 2 (the different tree selection criteria). This thesis shows that tree selection criteria applied to data transformed with a Hadamard conjugation can be placed in a statistical framework, and it is at step (5) that tree selection criteria with ties to standard statistical techniques appear to excel.

1.5.2 Statistical testing of phylogenetic hypotheses

We will outline the types of test most important to a phylogeneticist trying to reconstruct evolutionary history and estimate associated parameters. Later in this thesis we develop the mechanics for these and other tests, then apply them. We have already discussed the importance of Bayesian inference to synthesizing probabilities together so as to get a clearer picture of the overall situation. Here we look more closely at tests of specific details which can help refine our model and quantify our confidence in particular aspects of a tree.

For the construction of phylogenetic trees a primary test is "could my data have evolved by the process I am now using to model its evolution?". Of course we can reject this hypothesis in a more general sense, as we know that in sequence evolution there are complex biological and chemical processes which cannot yet be comprehensively modeled and are still hardly understood. Rather we want a measure of how close the data is to our models expectations, with a commonly used approach being to use a goodness-of-fit statistic such as X^2 (also known as the Pearson statistic, or rather imprecisely as the chi-square statistic). If the data is within the range we would expect for a random sample from the model, then we are comforted that other parameters of the model should also fall within the range of variability expected under the model.

There are also more specific tests of fit of data to model. For example amongst those models that fit acceptably we will also want to know if any fit appreciably better than others, so that we may give their parameter estimates more consideration than those of the other models (e.g.

Reeves 1992, Goldman 1993a). We will also need to make reliable decisions about how many parameters are important for the data to fit the model, compared to those that may be principally describing random fluctuations in the data. The key problem with increasing the number of parameters is that the variance of the most crucial estimates becomes larger. When adding too many parameters we also run the risk that under some circumstances the model will develop unpredictable behaviour (e.g. Stuart and Ord 1990, p.682). Likelihood ratio tests of nested hypotheses, or more general approaches such as the "Akaike information criterion" (or A.I.C., e.g. Sakamoto *et al.* 1986), aim to test which parameters in a model are useful and which can be discarded without expecting to loose overall accuracy.

Once in the region of reasonable parameter estimates, we can look at deriving their expected probability distributions under random sampling, in order to place confidence limits upon them. This may be done analytically, or increasingly with complex problems via computer simulations. Many simulations begin by using the bootstrap procedure for creating resampled data sets from the original data (e.g. Efron 1983, Efron and Gong 1984, Efron and Tibshirani 1986), with these bootstrap samples being called "fake data sets", bootstrap samples, or *pseudosamples* (the term we use throughout this thesis). With sequences this involves creating each pseudosample by randomly sampling with replacement a column of aligned sites a total of c times, where c is the sequence length (so in the pseudosamples some columns are missed entirely, others are sampled once, some are sampled twice, and so on)(Felsenstein 1985a, Penny and Hendy 1985). This is reasonable statistical procedure (e.g. Efron and Tibshirani 1986), but phylogeneticists must then apply a tree estimation method before obtaining estimates of the distribution of parameters, which hopefully reflect the stochastic error introduced by the original sample. Any non-linear nature in tree estimation can cause sample size biases in this procedure. In phylogenetics the method of bootstrap resampling followed by application of a tree building method, is often applied in order to estimate the proportion of times a particular internal edge appears in the replicate analyses (and each tree built from a pseudosample is called a bootstrap replicate or a *pseudoreplicate* of the original tree). The approach is equally valid to measure the sampling error of, for example, the length of an edge.

However if the method being used to reconstruct the data is not based on a reliable model we have no good reason to believe that bootstrap results are a reliable guide to the sampling distribution of an efficient and unbiased method (Felsenstein 1985). A secondary problem is assessing the type one and two errors associated with hypothesis testing via a bootstrap routine in phylogenetics (e.g. Zharkikh and Li 1992a, 1992b, Felsenstein and Kishino 1993, Rodrigo 1993). One possible solution to this last problem is the iterated bootstrap of Hall and Martin (1988), which Rodrigo (1993) applies and discusses in the context of phylogenetic tree estimation.

In terms of sampling distributions, the normal distribution (and its cohorts the chi-square, F, and Student's t-distribution) as elsewhere in statistics, are still the most widely used in phylogenetic statistical tests. This is because of various central limit theorems which often make them asymptotically exact, and also due to their mathematical tractability. They can be used in making useful tests of things like: "what are the expected confidence intervals on this edges

length?": "Is this parameter significantly different from that in another set of data?" (e.g. the number of transitions per site in one gene verses another). Multivariate tests under multivariate-normality are also generally quite tractable; for example, "is the sum of standardised errors for the fit of this weighted tree to this data within the range one would expect due to sampling error alone?" With this statistic one can then aim to construct a "confidence set" of all those trees that fail to be rejected by such a test. More complex tests such as a confidence limit for the ratio of one branch to another often require sensible (and robust) approximations since the true distribution is often not mathematically tractable (Stuart and Ord 1987, p. 325). In this thesis we describe a set of such tests, particularly for use with data analysed with Hadamard conjugations.

Another important challenge is evaluating the usefulness of asymptotic tests with sparse data. By sparseness we mean that there are many more possible patterns taken into account by the model than are actually observed in the data. For example the number of patterns possible grows as a exponential (e.g. 4^t where t is the number of DNA sequences being analysed), whereas sequences are finite in length. Further, the probability of observing a particular site pattern can vary by many orders of magnitude, making cell counts highly uneven. Sequence data is thus commonly both sparse and highly uneven in pattern probabilities, with the expectation that this will often make asymptotic tests quite unreliable (Read and Cressie 1988, Sokal and Rohlf 1981). Consequently, the chi-square approximation for the X^2 statistic of the fit of observed to expected sequence data can be unrealistic and misleading. Results in chapter 6 (and some results in Reeves 1992, and Goldman 1993a) suggest the problem is more severe than studies of other molecular data (e.g. Roff and Bentzen 1989) show. It is this sort of problem that phylogeneticists have to work around by using approaches such as simulations (Reeves 1992, Goldman 1993a) and resampling techniques (Rodrigo 1993). In chapter 6 we look at other possible solutions to this problem based on more accurate approximations, and grouping "like" cells into classes prior to application of test statistics (Read and Cressie 1988).

Another serious problem for statistical testing of phylogenetic analyses, is that many data sets generate new hypotheses which the researcher wants to test immediately. This is difficult because since the researcher had no prior hypothesis, it is difficult to know how many freak occurrences could have given something that looked as well supported as that observed? Rodrigo *et al.* (1994) have considered this problem with respect to interpreting bootstrapping of a data set claimed to give significant support for a grouping of metazoa and fungi. This factor of having no clear prior hypothesis alone places considerable doubt over many of the "statistical tests" currently being made in the field (something Felsenstein 1985 was careful to point out when introducing bootstrapping to phylogenetics). In this thesis we look at tests which will construct confidence sets of trees, in the hope that such an approach will allow corrections for hypothesis testing when the hypothesis is also generated by the analysis.

1.6 DESIRABLE CHARACTERISTICS OF TREE BUILDING METHODS

We now list and explain in a phylogenetic context useful criteria by which statistical methods may be assessed.

(1) CONSISTENT:

A consistent estimator is one which will converge to the population value (under a specified model) as the amount of data tends to infinity (here as c, the length of sequences, increases). For biologists it is often the unweighted tree that is of most importance. As such, if a tree building method obtains the unweighted tree with long sequences, it is called consistent. An all or nothing estimator (like the unweighted tree) may give the correct answer in some parts of the parameter range, but the wrong answer for other parameter combinations. Consequently an estimator of the unweighted tree may be called both consistent and inconsistent for the same mechanism of evolution, depending on which part of the parameter space it is applied to. This usage agrees with that of Felsenstein (1978), in the seminal paper on the consistency of tree reconstruction methods.

Many tree building methods aim to recover weighted trees, in which case the most strict definition of consistent is that under a specified model methods must recover the tree and edge weights exactly. Once the data are aligned (1), correcting for multiple hits (2) comes next. For a distance method (and Hadamard conjugations) the data are corrected for multiple changes independently of identifying any tree. To recover correct edge weights this step must be consistent i.e. converge exactly to the true edge weights (the values associated with the tree that generated the data). Once a tree selection criterion is chosen (3), the next step (4) is an algorithm to choose the true tree from either its path lengths, or in the case of the Hadamard conjugation, its indexed edge lengths. How well the algorithm at step (4) performs is often dependent upon the effectiveness of compensating for multiple changes in step (2)(becoming more dependent as the number of multiple hits increases). In the case of building trees from distance matrices many consistent tree identification methods are known given that they are provided with tree additive data (however see later in this chapter where we require a tighter definition of additive than usually used, since some transformations can sometimes produce additive distances on an incorrect tree). As an example, since the match of distances to tree must be exact, then any method which minimises a monotonic function of mismatch must also be consistent, i.e. will identify the correct tree (e.g. the weighted least squares fit of Fitch and Margoliash 1967, the minimum evolution criterion of Kidd and Sgaramella-Zonta 1971, and Rzhetsky and Nei 1992a). This of course does not guarantee that a tree evaluation criterion will not be mislead some of the time by sampling errors due to finite sequence length. Some methods of tree selection which combine steps 3 and 4 together have recently been identified as inconsistent with additive distances e.g. Li's (1981) method, which is noted, with a correction offered, in Charleston *et al.* (1994). Fortunately most tree selection algorithms are guaranteed reliable with additive distances.

With two useful definitions for the consistency of a tree reconstruction method we need to be specific about which one we are using. However for estimators of continuous variables (e.g. a distance) then there is just one definition, convergence of the estimator to the exact value. Likewise a Hadamard conjugation returns values for the length of possible edges in the tree that generated the data (a continuous variable). All edges not in the tree generating the data have an expected value of zero, if this is not the case with infinite data, then we should call the Hadamard conjugation inconsistent. Inconsistency in itself is not so crucial here; a tree selection criterion applied to data transformed by an inconsistent estimator may still have a better chance of reconstructing the correct unweighted tree, and probably also making better estimates of edge lengths, than without the transformation. It is also very obvious that practically all our estimates of continuous valued quantities from real data must also be inconsistent (not the least because we do not know the underlying mechanism of evolution for a particular case). What is more important is the size of the errors being made, or in the case of tree building methods, how often a method gets the correct unweighted tree for a finite amount of data. So consistency of continuous variables is most often just a useful term to tell us when a method is exact under a certain model; it is the size and consequences of errors that we should really be focusing upon.

(2) STATISTICALLY EFFICIENT (MINIMUM VARIANCE + UNBIASED)
-RAPID CONVERGENCE TO THE TRUE TREE

A very desirable attribute of a statistic is that its variance about the true value be minimal for a certain amount of data. If a statistic has minimum variance for any amount of data it is called efficient (actually most estimators are only known to be asymptotically efficient, that is, after some unspecified amount of data they are guaranteed have lower variance than all others). A statistic is *biased* if for some amount of data (e.g. sequence length) the mean of the statistic, $\bar{x}$, does not match the true value, μ (even if it does converge to μ for larger amounts of data). The variance of a sample statistic about its true value (also called the mean square error) can then be divided into two components $V_\mu(\bar{x}) = V(\bar{x}) + (\bar{x} - \mu)^2$, where $V(\bar{x})$ is the sampling error about the sample mean, and $(\bar{x} - \mu)^2$ is its bias (the difference between population mean and sample mean, which can be considered systematic error) (Stuart and Ord 1990, p629). Generally we desire an estimator that will minimise both components together (for example a minimum mean square error estimator).

A good example of bias generation is that which occurs when correcting short sequences for multiple changes. Most such methods take *natural logarithms* (*ln*) of some observed quantities. However $E[\ln(x)] \neq \ln(E[x])$, and the difference in these values is the bias which in this case goes to zero as $\text{var}(x) \rightarrow 0$. Defining bias in a tree building method will depend upon how we measure the variance of its errors. If we use a continuous variable such as the distance in tree space between two trees, or from tree to data (e.g. Robinson and Foulds 1979 on the continuous partition metric, Hendy 1989, Steel and Penny 1993), then we can define sampling variance and bias in the traditional sense. If we are to use the recovered tree as a discrete estimator, then Kuhner and Felsenstein (1994) offer a definition of bias in this case (note that these authors also reinvented the Robinson and Fould's 1979 metric for this purpose). The important thing is that

the accuracy of tree recovery is a result of both factors, that is variance of estimates about their sample mean, and the separation of this sampling mean from the true tree, the bias.

(3) ROBUST:

A robust estimator is one whose accuracy is least seriously compromised as we violate its assumptions, for example, change the parameters and or mechanism away from those under which the estimator was defined to be consistent. While the mechanisms of real sequence evolution are unknown, even the present generation of models have been little studied from the perspective of robustness (one exception is found in Hasegawa and Fujiwara 1993). Two biological processes which may well lead to inconsistency of tree building methods applied to real sequences are, non-stationary substitution transition matrices (possibly due to both mutation and selection) and variation of rates of evolution across sites (usually expected to be due to stabilising selection, and or a covarion model, see Fitch and Markowitz 1970). At present it is not possible to model these processes accurately, so caution is called for in interpreting any analysis of ancient molecules where such process may have fluctuated massively (see Lockhart *et al.* 1995, for a discussion of such factors in the evolution of genes associated with photosynthesis).

The robustness that biologists need to know about is a function of methods and the true process of evolution; a thorough evaluation will demand a greater understanding of the evolution of real molecules. Evaluations in this thesis aim to gain further empirical and theoretical understanding of the relative robustness of a variety of methods.

(4) TESTABLE , POWERFUL AND INFORMATIVE:

We wish to have well categorised statistical tests of our phylogenetic methods. We have already discussed what we mean by *testable*, and mentioned some types of test which should be made. We put special emphasis upon the test of fit of data to model, as the first crucial test. Such a test can serve as a guide to how we may best edit our data to get the maximum separation of phylogenetic signal from extraneous influences, and as a pointer to possibly better models. By *powerful*, we mean a test which is sensitive to detecting departures from the null model (note that this is the statistical use of the term, some authors, e.g. Penny *et al.* 1993 refer to a tree selection being powerful when they mean it is statistically efficient and thus has rapid convergence). Strictly speaking, the *power* of a test is the probability of making what is called a type two error in hypothesis testing, that is failing to reject the null model when it is indeed wrong. Testability is a necessary prelude to being able to adjust the probability of a hypothesis being correct after an analysis is performed.

The number and diversity of tests that can be performed is a measure of how *informative* a method will be about the true mode of evolution (e.g. distribution of rates across sites, transition to transversion, indications on which sites are most likely not following the model imposed). More *powerful* statistics (e.g. tighter confidence intervals) make for more *sensitive* methods, which hopefully will require less data to reach significant conclusions about the nature of evolution. The methods which presently make the most intensive use of the data, especially maximum likelihood and Hadamard conjugations, are also the methods which appear to be the

most precise and informative. By giving quantitative and well understood statistics of fit, methods allow quantitative assessment of phylogenetic results and thus make many hypotheses testable. In this thesis we evaluate how sensitive different methods are to detecting violation of the model they are based upon, and in chapter 5 we note a clear advantage in this area for sequence based methods over distance methods.

(5) COMPUTATIONALLY EFFICIENT:

Can we do the necessary calculations on a given set of data? A useful way of defining the efficiency of calculations is by the concept of the order of computational efficiency (called simply efficiency in mathematical usage). There are two sets of calculations involved in finding an optimal tree; assessment of any given tree, and searching across the landscape of all possible trees. We will define the order per tree as the cost of evaluating each tree. Ideally a method is O(n) (order n), so that the time to evaluate a larger data set increases linearly with the number of taxa included in the analysis. Orders n^2 to n^4 are also promising in that increases in computer speed will translate to more sequences being able to be analysed all together. The order of operations is, however, a measure of worst case performance, and tells us little of how much better a method might be expected to do with well structured data, as sequence data often is. Unfortunately there is no reliable measure for the computational efficiency of searches amongst trees for "typical" sequence data, as both sequence length and the number of taxa are varied. Well thought out simulations should be able to give some reasonable guides to how different methods perform.

While methods such as maximum likelihood are generally considered computationally expensive, it is also worth remembering that the calculations being made also yield many valuable ancillary statistics (e.g. confidence intervals on parameters) which can be expensive or even impossible to calculate with other methods. In the case of maximum likelihood little work has been done on the order of its calculations, nor is their any indication of how quickly the cost of these evaluations rise as more sequences are added. Practically speaking, the bottom line in computational efficiency for a method proposed for wide use in phylogenetics, is whether the necessary calculations can be done in a reasonable period of time (typically no more than a weekend or two). If two methods can be completed within an acceptable time frame, then the choice between them is not which one took the fewest hours, but rather which is yielding the most useful results (which must include the ability to make statistical tests).

1.7 CURRENT METHODS OF TREE ESTIMATION

Previous attempts have been made to classify tree building methods into discrete classes (e.g. Swofford and Olsen 1990, Penny et al 1992). However due to improved understanding of the different methods, some of their properties need to be revised and reconsidered. Here we describe the methods and consider some of their respective advantages and weaknesses.

1.7.1 Maximum likelihood on sequences

Maximum likelihood (ML) is a statistical criterion for selecting parameters values in a model. It does so by choosing those parameters which are most likely to generate the observed data. If we assume independence between sites, then observed sequence data is considered to constitute a multinomial sample. To calculate the likelihood of the data we need the probabilities of the observed sequence patterns (here called s_i) under a model. Given these probabilities and the assumption of independent sites, the natural logarithm of the likelihood is just the sum over all patterns of $f_i(obs) \times \ln(s_i)$(where $f_i(obs)$ is the observed frequency of the *i*-th pattern). The first requirement is to calculate the probabilities of the data under a specific model in a reasonable period of time given available computing power. Next we must search for the optimal set of parameter values, and depending on the parameterisation of the model may need to evaluate many parameter sets (e.g. different unweighted trees).

The application of ML in phylogenetics has been pioneered by Felsenstein (1973, 1981a). It requires the development a method of calculating the probability of a particular sequence pattern, under a particular model which includes specifying the weighted tree. The likelihood, typically measured as the natural logarithm lnL, is then iteratively optimised. This method has been applied for up to 12 parameters per edge i.i.d. models (Barry and Hartigan 1987). More recently it has been shown how the probability of sequence data can be calculated when there is a distribution of rates across sites. In Steel *et al.* (1993c) and Yang (1993) this is done by assuming a defined form to the distribution. In the method of Felsenstein and Churchill (discussed in Felsenstein 1993) the approach is to define different discrete rate classes, their relative probabilities, how long runs of a given rate class are expected to be, and then use hidden Markov chains to infer the likelihood of the data. Unfortunately the computational cost of this type of ML model can be much larger than i.i.d. methods, and often requires very expensive numerical integrations. There is some latitude for methods to be sped up with optimisation of code to particular machines, and Fast DNAML (Olsen et al 1992) is an example allowing ML tree selection to be applied to 50 + sequences on supercomputers. The other hope is to find sensible approximations which hopefully will have minimal impact upon reliability (one example is how small decreases in likelihood on a tree need to be before stopping an iterative optimisation). Given that the proportion of possible trees searched by any method must rapidly decrease as the number of taxa increases above say 25, it remains unclear how reliable different combinations of tree selection criterion and search procedure are. It is not inconceivable that likelihood evaluating 10,000 different trees will identify a tree closer to the true tree than say parsimony evaluating 1,000,000 trees. A possible reason for this being true could be that as long as you are searching

in the right neighborhood, it is the ability to reliably discern amongst similar trees that makes all the difference to success or failure.

The principle of ML is generally well understood, and is a highly favoured method of statistical analysis, not only for point estimation but also in terms of understanding the whole model-data relationship (Edwards 1992). The variance-covariance matrix of the model provides a useful summary of the likelihood surface about a given tree, and allows useful statistical tests to be made (e.g. see Felsenstein 1981, 1982). All the variables in the model are continuous and also correlated, so there must be some concern about how good implementations are at finding a global optima for a given tree. Both flat likelihood surfaces and distinct positive or negative correlations between parameters (e.g. edges as shown in Waddell *et al.* 1994, and in chapter 3 of this thesis) can cause search procedures to prematurely terminate (discussed in this thesis). So to can unevenness of the likelihood surface, and while Felsenstein (1981a) pointed out that asymptotically (with sequence length $\rightarrow \infty$) under an i.i.d. model the likelihood surface becomes multivariate normal, it is unknown how quickly it gets there. Newton and quasi-Newton methods are generally faster to converge than single line searches (e.g. as used by the latest versions of PHYLIP) or the E-M algorithm (Stuart and Ord 1990, p. 694-695) which was used in the earlier versions of PHYLIP. Overall however, little is known of the characteristics and relative merits of different numerical methods for application to different sized trees and phylogenetic data. Note that the entries of the variance-covariance matrix are equal to minus twice the inverse of corresponding entries in the Hessian (second order derivative, or information matrix) and this provides one way to calculate them.

A further concern for the method of ML is that there may be multiple optima per tree, something which Steel (1994), has recently illustrated (using a four taxa tree and two character patterns). More recently (M.A.. Steel pers comm.) has demonstrated that data sets with many taxa and multiple sequence patterns can have the same problem. Although all these examples still seem rather specific, they must remain of some concern until (if) it can be shown that this problem becomes insignificant for typical sequence data sets. If not then numerical methods may need to find multiple optima by starting at different points in the landscape, something which could dramatically increase necessary computations.

An alternative way of looking at ML methods is that they aim to optimise the "fit" between data and model, a view which is emphasised in this thesis. This fit is often measured using the G^2, likelihood ratio statistic (e.g. Stuart and Ord 1990, chapter 30) between the observed pattern frequencies and those predicted by the model. A close relative of ML methods are minimum chi-square methods (e.g. Pearson's X^2 statistic, Stuart and Ord 1990, chapter 30). Chapters 5 and 6 look at the relative power and robustness of tree building methods based upon each criteria.

1.7.2 Parsimony, compatibility and closest tree

Parsimony and compatibility are well known tree selection criteria which do not themselves make any correction to the data for multiple changes. When applied to observed data, these tree selection criteria can converge to the wrong tree as more and more data is added (i.e. they are

inconsistent in selecting the unweighted tree) as pointed out by Felsenstein (1978). However this same problem besets any tree selection method (a combination of optimality criteria and correction for implied changes) that makes insufficient adjustment for parallel and convergent substitutions. Consequently (Steel et al 1993b) pointed out that parsimony, compatibility or closest tree based tree selection procedures are consistent if the data originated from a model for which Hadamard conjugations are consistent, and this transformation is applied to the data prior to tree selection. Understandably readers may feel surprised at this conclusion, mostly because parsimony has historically been closely associated with researchers who have worked with morphological data. Such data has evolved by a complicated mechanism, and there has been no general mechanism proposed. Yet even so corrections to the data are often practiced (e.g. reweighting of characters, one of the most extreme of which is which characters to admit to an analysis). Accepting this, it is easier to see the validity of applying parsimony to the non-linear reweighting of characters which, to a large extent, do share a common underlying mechanism of change. Both parsimony and compatibility require slight modification so that they can deal with data giving some patterns negative values. At present this is done by ignoring such patterns, although other possibilities remain to be explored.

Felsenstein (1981b) made the interesting observation that parsimony and compatibility are interconnected by a series of intermediate methods which provide increasing maximum penalties (going from compatibility to parsimony) for character patterns which do not fit the tree under consideration (a type of dynamic weighting). Compatibility is analogous to minimising the sum of deviations (ignoring negative deviations), while parsimony minimises the sum of tree weighted deviations. Interestingly Charleston (1994), has recently shown that parsimony significantly out performs compatibility in tree selection when rates of change are relatively low, which may well be due to its weighting effect speeding up convergence (see also Felsenstein 1981b for related discussions). Both methods can have their robustness and convergence properties enhanced under many models (especially when rates across sites significantly differ) by appropriate site by site weighting of the data (e.g. Penny and Hendy 1985, 1986, Williams and Fitch 1989, Hillis *et al.* 1994). Unfortunately there are few studies of the robustness of these methods, which certainly do not guarantee consistency (as discussed in the exchange of Chippindale and Wiens 1994 with Huelsenbeck et al 1994).

Lastly, Hendy (1989) introduced a tree selection method called "closest tree." What this does is select the weighted tree closest in Euclidean space to a vector of sequence patterns (relying particularly upon sequence bipartitions). As we will describe later in chapter 5, closest tree is very similar to taking only character state bipartitions, squaring the value of each one, then picking the largest clique (i.e. using compatibility) amongst the reweighted patterns. As such, it is very similar to an ordinary least squares tree selection procedure. Unfortunately results studying the effect of sampling errors on the Hadamard conjugation in chapter 5, suggest that "closest tree" does not have any particularly desirable statistical properties when compared to other criteria such as weighted least squares, parsimony or compatibility (while simulations in Charleston *et al.* 1994 show it having inferior performance to both compatibility and parsimony with Cavender model data, transformed by the Hadamard conjugation).

1.7.3 Distance based methods

Distance based methods have the advantage that there are many formulae known for making pairwise distances additive in expectation under a variety of models (see for example Zharkikh 1994). When using nucleotide data, sequences are first reduced to a matrix of pairwise dinucleotide counts, for example how often does an A align to an A in the other sequence, an A to a C etc. Next a non-linear transformation is applied with the aim of making the distances closer to additive by estimating what the total distance would be if unseen substitutions are inferred and included. Additive path correction formulae are now available for all i.r. and i.i.d. stationary time reversible models (Lake 1994, Steel 1994a), and for all time reversible i.i.d. models if we know the distribution of rates across sites (chapter 3).

Barry and Hartigan (1987a) proposed "asymmetric distances" based on logarithms of determinants, which they claimed would give tree additive distances under any i.i.d. model. The distance returned by this formula is not the usual expected number of substitutions per site, but rather a weighted number of substitutions per site (Barry and Hartigan 1987a). However their distance is asymmetric under nonstationary base composition models, and in chapter 3 it is shown this feature can lead to inconsistent tree reconstruction. Other early applications of the utility of the logarithm of the determinant of Markov transition matrices on trees include Cavender and Felsenstein (1987) and Rodriguez and Medina (1986, unpublished). More recently log determinant distances have been defined in a manner which is guaranteed to give tree additive distances under any i.i.d. model, and they have been applied to biological problems where they apparently give more reasonable results than standard distance methods (Steel 1994a, Lake 1994 and Lockhart *et al.* 1994). In this thesis we look at ways making log determinant distances robust to a distribution of rates across sites; a factor which any method must cope with if it is to be routinely used for analysing coding sequences. Using real data we take a critical look at how useful log determinant methods really are by looking at their sampling variance, bias and the stability of a tree selection criterion working from them, in comparison with commonly used distance measures.

Approximately a dozen consistent algorithms have been published for recovering trees from additive distance matrices. Most of those using explicit tree selection criteria are based on least squares fit (see Felsenstein 1982, 1988, and 1993 for reviews, except for generalised least squares discussed by Hasegawa *et al.* 1985 and Bulmer 1991) or else the minimum evolution criterion. Others are algorithms combining optimality criteria and localised tree searching together (e.g. neighbor-joining, distance Wagner) as discussed already. Some of the distance based methods (especially generalised least squares, GLS) allow a set of statistical tests, including fit of data to model (Hasegawa *et al.* 1985, Bulmer 1991). As we show in this thesis an iterated GLS method can be considered the maximum likelihood estimator when using just pairwise distance data. (It is interesting to note that metric distances do not guarantee reliable tree selection, while additive distances will e.g. Felsenstein 1984). Note that when using the term metric it is important not to confuse the usual definition of metric given in section 1.3, with Buneman's four point metric which is a direct consequence of additivity. Buneman's four point

metric simply claims that on a tree of four taxa, then with additive distances, $\delta_{ij} + \delta_{kl}$, will be minimal if the tree is (i,j), (k,l), in which case both $\delta_{ik} + \delta_{jl}$ and $\delta_{il} + \delta_{jk}$ will be larger and equal.

Lastly in order to make the claim that additivity of distances on a tree guarantees consistency we must define a distance measure's additivity as existing upon all possible weighted trees which have the tree of interest as a subtree (and in this sense we will use the term "additivity" generally throughout the thesis). This qualification is necessary since we have found some combinations of data and transformation which result in additivity of transformed distances upon "incorrect trees" (see section 5.5.5). The qualification simply states that in order to be called "tree additive", a distance must be additive upon all possible trees, with all possible edge weights.

1.7.4 Phylogenetic invariants

A tree invariant of a model is simply a quantity that tends to a constant value dependent only upon the unweighted tree (see Steel *et al.* 1993c for a summary of such concepts). With finite data, invariants can provide evidence (subject of course to statistical fluctuations) as to which edges were in the tree that generated the data. Three different types of tree invariants are presented by Cavender and Felsenstein (1987), Lake (1987), and Hendy *et al.* (1987). Probably the simplest are the linear invariants of Lake (1987) which he dubbed evolutionary parsimony. These are three linear functions of observed sequence site patterns that asymptotically identify the unrooted tree for four taxa. This method aims to identify the branching pattern, but not the edge lengths. One of these invariants has an expected value of zero, while the other two will be positive and equal. A number of authors have shown that these invariants each have a binomial distribution, being sums of cells from a multinomial distribution (e.g. Navidi *et al.* 1991). This allows a statistical test of whether one tree is significantly better supported than the other two and equally importantly, whether the values of the invariants are within the range expected by the model (a test of fit of data to model). A useful consequence of the calculations being linear in the observed data is that this method has the desirable property that it remains consistent when rates vary across sites. However all sites must retain the same relative substitution rates throughout the tree.

While Lake's (1987) invariants allow sites to have a distribution of rates across sites, relatively fast evolving sites will slow down the rate of convergence by increasing the variance of all three invariants. Because each invariant uses only a subset of the 256 site patterns (actually 12 of the 36 unique patterns under the Kimura 2ST model) to evaluate each tree, the method suffers from low statistical efficiency i.e. slow convergence (Jin and Nei 1990, Navidi *et al.* 1991). (See Li *et al.* 1987 for a review of Lake's method). Until now this method has been limited to analysing four taxa at once. Accordingly, in order to build a tree for more taxa these four taxon subtrees must be combined, which can be done in a variety of ways. (Note though, that Lake's original usage was to evaluate support for a single hypothesised internal edge by taking all combinations of one taxa from each of the four groups connected by that putative edge). Recent developments, prompted in part by findings in this thesis, have resulted in the identification of linear tree invariants under the generalised Kimura 2ST model and also the Jukes-Cantor model.

With five or more taxa, some of these new will take an expected value of zero only when the true tree has two or more specific internal edges in it. These developments, plus recent related work by Fu and Steel (in press), and Hendy and Penny (in press), are explained in chapter 4.

Other types of invariants (called model invariants) also exist when the data comes from a certain model. A special case of these are clock invariants, which include the constraints imposed by ultrametric distances (the basis of most relative rates tests). Tajima (1993b) lists some such invariants. The polynomial invariants of Cavender and Felsenstein (1987) exist under two state models, but have rarely been applied. Many other tree and model invariants have been found under a fairly wide variety of mechanisms of evolution, but to date none of these have been shown to be particularly useful for either tree selection or model evaluation. Continued interest in phylogenetic invariants rests largely upon the hope that they will show up specific properties of models which may lead to faster algorithms for ML in particular.

1.7.5 Invariants by invertible transformation of sequences

A more extensive class of invariants is that introduced by Hendy *et al.* (1987) for two state characters (e.g. purines/pyrimidines), and both extended and streamlined in Hendy and Penny (1993). Since then it has been extended to 4-state nucleotide sequences by Szekely et al (1993), Steel et al (1992), Steel et al (1993c) with a first application in Hendy et al (1994). This method uses all the data for the model it is based upon, and is fully invertible. It can go from sequences to a vector description of the tree that generated the data (with all possible edges that are not on the true tree having value zero, and positive entries corresponding to the edge weights in the true tree). Alternatively this method can start with a weighted tree and generate a vector of the probability of every sequence pattern under that model. Within this thesis the edge weights associated with the method (known as a Hadamard conjugation) are always measured as the expected number of substitutions per site.

The method of Hendy and Penny (1993) has a fascinatingly simple structure. First an ordered vector of all sequence pattern frequencies (called $\mathbf{s}$) is multiplied by a Hadamard matrix ($\mathbf{H}$), which because of its form corresponds to a discrete Fourier transform. The result of this operation is a calculation of the observed length of all possible pathsets for the taxa being analysed (vector $\mathbf{r}$). Next a correction for multiple changes is applied without needing any knowledge of the true tree (just like a distance correction) in order to infer additive pathset lengths (vector ρ). Lastly the vector of corrected generalised distances is multiplied by the inverse of the first Hadamard matrix ($\mathbf{H}^{-1}$). This recovers quantities analogous to site pattern frequencies, but corrected for multiple changes (contained in vector γ). (In fact if the data is without sampling error and generated under the model, then the only entries greater than zero will be the weight of each edge in the tree generating the data). Of great utility to the application of the method is the existence of a discrete Fast Fourier transform (the Fast Hadamard) that requires far fewer operations than the usual method of matrix multiplication (see Tolimieri *et al.* 1989, and Hendy and Penny 1993). This is purely a computational aid and does not alter the

statistical structure of the method so is rarely referred to directly in this thesis, but it is of great assistance to the programming and application of the method to larger data sets.

The term Hadamard conjugation is an apt description of the method because a conjugation refers to a mathematical operation (multiplication by **H**), followed by a second operation (a nonlinear transform to correct pathset lengths for multiple changes), followed by the inverse of the first operation (multiplication by $\mathbf{H}^{-1}$). The method is mathematically invertible because the pathlength correction is a one to one monotonic function with a known inverse. Because the result of a Fourier transform is often called a spectrum, then this method is also called *spectral analysis* by Hendy and Penny (1993). We prefer not to use this term as it has been used by quite a few authors over the past decade to describe other phylogenetic techniques (some but not all being other types of invariants).

A relative of Hadamard conjugations, the distance Hadamard (also described in Hendy and Penny 1993), starts with just pairwise distances and infers corrected pathset lengths from the minimum sum of sets of corrected pairwise distances. These pairwise distances can be calculated in any way and do not require a prior multiplication by **H**. In collapsing data from sequences into pairwise distances, much information is lost (Penny 1982, Steel *et al.* 1988) and even with the two state equifrequncy model we cannot reconstruct the original sequence pattern probabilities when using more than 3 taxa. Informative comparisons between the sequence Hadamard conjugation, and the distance Hadamard operation in chapters 4 and 5, help to reveal some of the innate differences between these procedures.

The next step after a Hadamard conjugation is to choose a tree from the vector of inferred edge weights and invariants (γ) (entries with weight zero, not corresponding to any edge in the tree). If the data fits the model, then after the Hadamard conjugation the only positive entries correspond to edges, so selecting the tree is straightforward with large amounts of data. With finite sequences an important question is: "Which tree selection criterion will work best given both sampling error and systematic error in (γ) when the model is violated?" In chapter 4 we look at how sampling error causes random fluctuations in γ, specifically showing how to calculate their combined variance-covariance matrix and looking at the form of sampling distributions. In chapter 5 we are then able to consider further the merits of statistically based tree selection criteria which go beyond methods such as parsimony, compatibility and closest tree.

1.7.6 Other ways to classify tree estimation methods

In summary then we can divide methods of tree building up into those which work directly with sequence patterns (sequence ML, Hadamard conjugation and tree selection algorithms such as compatibility and parsimony applied to observed sequences) verses methods which work from a matrix of pairwise distances (neighbor joining etc.). Penny (1982) pointed out the vast difference in information content of these two types of data as the number of taxa increases and speculated that this factor must be having a substantial effect upon the performance of different methods. To date, few such effects have been detected let alone diagnosed. In chapter 5 we

detect some such differences between otherwise matched methods, and show they can be substantial with as few as four taxa in the data set.

Having described the main steps and methods used in tree estimation, we can now describe another enlightening way of grouping tree building methods. This is based upon the way different method(s) attempt to overcome the effect of parallel and convergent changes misleading the tree building method. The main classes of "data correction" presently being used are as follows:

(1) No alteration to the data, other than editing or conversion to pairwise distances. This is the method of choice if one feels that the data had few parallelisms and convergences, so that either multiple substitutions at a site were having no substantial effect, or any attempt to "correct" the data would result in worse systematic biases or increased variances (i.e. magnified stochastic errors). The second point being a real concern with very short sequences of fewer than 100 sites.

(2) Data transformation based on a mechanism of substitution, such as the Kimura 2ST distances, which aim to make the data more additive (a special case of generalised linear models as already mentioned). Reduction to distance matrices, followed by non-linear transforms are the best known approach, while Hadamard conjugations are a recent alternative.

(3) Maximum likelihood methods for sequences (e.g. Felsenstein 1981a, Barry and Hartigan 1987a), which aim to express the effects of parameters in the same form as the original data and optimise free parameters by "fit" between observed and expected data. Minimum chi-square methods (X^2) are close relatives, which can involve nearly identical calculations. Linear invariants can also be viewed as special case of this approach (Navidi *et al.* 1991).

(4) Approximate site weighting schemes. These may be applied to the observed data with or without some iterative estimate of phylogeny (e.g. Farris 1969, vs Hendy and Penny 1985), or they may be integrated in with any of the previous methods (e.g. Olsen 1994's site weighting ML approach).

(5). Approximate substitution weighting schemes. Such methods so far have been most closely associated with parsimony tree selection applied to the observed data (e.g. Sankoff 1975, Sankoff *et al.* 1976, Williams and Fitch 1989), but could be extended to any of the other methods (for example distances, Schöniger and von Haeseler 1993).

Some form of corrections will hopefully have the multiple effect of making the tree building method consistent with the data being analysed (in terms of the unweighted tree), less biased (in terms of edge lengths), and give rise to more statistically efficient tree selection. We often desire that the method of making "corrections" will be model based so as to allow a whole battery of statistical tests for studying sequence evolution. Of these five classes of corrections, only the second and third are explicitly model based and offer the best possibilities for powerful statistical tests. In this thesis we study properties of the first three types of "correction" in relation to one another.

1.8 MAIN RESULTS IN THIS THESIS

Work in this thesis particularly centers around a statistical understanding of Hadamard conjugations, and how they might best be used as a reliable and versatile scientific tool. Presently its utility is largely confined to that of being point estimators with no defined statistical structure. This makes it difficult to effectively apply Hadamard conjugations in many practical applications, especially where sampling errors are large. Fortunately the conjugations mathematical structure makes them excellent candidates for the development of a statistical framework.

Important new results of this thesis include to:

* Extend Hadamard conjugations to incorporate the relative rate of substitution varying with the site (the unequal rates across sites effect or URAS).
* Modify time reversible distances and the LogDet transform to accommodate a distribution of rates across sites.
* Study the mathematical and statistical properties of the LogDet class of distance corrections.
* Examine different ways of optimising parameters associated with a distribution of rates across sites.
* Derive the variance covariance structure, and study the sampling distribution and bias of both Hadamard conjugations and the distance Hadamard method.
* Show that amalgamating then averaging equivalent sequence patterns of the Kimura 2P and Jukes-Cantor models restrict the 4-state Hadamard conjugation to these models (thus reducing sampling variance in the transformed data).
* Consider how to deal with samples which give negative entries in the Hadamard intermediate vector **r**, and hence make the logarithmic correction inapplicable (unless one considers complex numbers!).
* Show that a recently derived formula for estimating pairwise distances (that of Tajima 1993a) is extendible to Hadamard conjugations. We also show that this method is not so important as a reduced bias method, but rather as a reduced mean square error method in important instances.
* Describe statistically based methods of selecting trees from Hadamard conjugation transformed data.
* Describe some relationships between ML, and tree selection after a Hadamard conjugation.
* Evaluate the robustness of tree selection criteria in the well known Felsenstein zone.
* Describe a problem we name the inverse or anti-Felsenstein zone and examine a "long edges repel" problem, which can occur when data are over corrected for multiple changes. Sequences with nonstationary base composition may be quite vulnerable to this effect.
* Evaluate the severity of the anti-Felsenstein zone for a variety of different tree selection procedures.

* Evaluate convergence properties of tree selection on observed data, and Hadamard conjugation or distance Hadamard corrected data.

* Implement, evaluate and discuss ML with variability of rates across sites, including continuous distributions (the gamma distribution (Γ), and the inverse Gaussian distribution), plus mixtures of these distributions with invariant sites.

* Consider ML models where the relative rates at sites are non-stationary or where the phylogeny is reticulate.

* Discuss the utility of branch and bound to find the optimal ML tree.

* Define and discuss tests of the fit of data to model.

* Consider the meaning of "a confidence set of trees".

* Describe tests and confidence intervals for edge lengths, tr/tv ratios, and distributions of rates across sites for Hadamard transformed data.

* Look at the error structure involved in using evolutionary trees to estimate sequence divergence dates and hence infer rates of molecular evolution. We then bring in the additional errors associated with coalescent times if we wish to estimate species divergence dates.

* Illustrate new techniques with DNA sequences relevant to two of the most interesting questions in phylogenetics, namely very early evolution and human-ape relationships. Results include:

(1) A high degree of support for humans and chimps being closest relatives, even when third position mtDNA sites are retained in the analysis

(2) A most probable divergence data of human and chimp lineages less than 6.5 million years ago with a standard error of approximately 1 million years.

(3) A discordance between the results from the data set / alignment combination of Lake (1988) vs Gouy and Li (1989a). Alignment errors may be deceiving an ML analysis of a subset of four taxa to attribute a good fit of data to model, and suggest significant support for an incorrect tree.

(4) A reanalysis of Gouy and Li's (1989a) alignment of 16S-like rRNA from 28 taxa across the tree of life using LogDet and other methods of pairwise distance correction made robust to variation of rates across sites suggests:

-That the earliest divergence amongst living eukaryotes involved the separation of the Microsporidia and not the diplomonad (e.g. *Giardia*) lineage (in contradiction to practically all molecular analyses to date).

-Support for archaebacteria being monophyletic which is revealed by the most conserved characters, but may be disguised when assuming i.i.d. models.

-That while many other features are in good agreement with previous analyses, there appears to be strong evidence for the monophyly of plants and animals (exclusive of fungi). This result deserves close scrutiny, given claims to the contrary based on other analyses of the same type of molecule (e.g. Wainright *et al.* 1993, Baldauf and Palmer 1993).

1.9 DATA SETS ANALYSED IN THIS THESIS

Because we are developing new methods, it is desirable to illustrate what is happening in the various data transformations and tests. New methods in this thesis are primarily illustrated with a four taxon and a six taxon data set, although in chapter 3 we use a 28 taxon data set extensively.

1.9.1 A subset of Lake's alignment of rRNA molecules

The first data set is relevant to understanding the relationships of Archaebacteria, Eukaryotes and Eubacteria. The sequences are from the more slowly evolving conserved regions of the small subunit 16S-like rRNA. One set of sequences is from Lake's (1988) alignment and includes human (*Homo sapiens*), a eubacterium (*Escherichia coli*), and representatives of two main groups within archaebacteria, an acidophilic thermophile (*Sulfolobus solfataricus*) and a halobacterium (*Halobacterium volcanii*). Lake's (1988) alignment is included as one of the example data sets with the program MacClade 3.0 (Maddison and Maddison 1992). A similar set of sequences (substituting *H. volcanii* for its close relative *H. salinarium*) has been used by others developing new methods of analysis (e.g. Navidi *et al.* 1991, Churchill *et al.* 1992) and allows some useful comparisons. The analysis of this data is controversial with some claiming it supports a closer relationships of some extremely thermophilic and acidophilic Archaebacteria with Eukaryotes, to the exclusion of many other archaebacteria (e.g. Lake 1988). In chapter 2 the effect of variation of rates across sites seems crucial to deciding on the best tree for this data. Lake (1988, 1989) considers this data an exemplar of an unequal rates tree like that in Felsenstein's (1978) paper. Because the period of time since the divergence of the taxa is so long (probably close to 3 billion years) then:

(1) relative rates rate of substitution in each lineage could be quite different, leading to a tree like that of Felsenstein (1978).

(2) the underlying processes of nucleotide mutation affecting substitution patterns could strongly diverge, resulting in non-stationary relative probabilities of different types of substitution (e.g. independently evolved GC biases resulting in more convergent nucleotide substitutions than the model predicts).

(3) For sequences to still be informative after so much time there must be some form of strong stabilising selection. Strong functional constraints on the molecule could dramatically alter the rate of substitution at different sites. In addition the relative rates at different sites could also be changing through time, perhaps due to a covarion type model (Fitch and Markowitz 1970).

(4) Methods used to align such sequences typically require a set input order, resulting in a bias towards more matches for the sequences aligned first (e.g. Lake 1991). To counter this effect Lake (1988) aligned all sequences against the sequence which appeared to be on average the least diverged from all others. However it is suspected that his alignment is in fact quite flawed (Olsen and Woese 1989). Lake's critics (e.g. Gouy and Li 1989a, Olsen and Woese 1989) were in

contrast also careful to exclude difficult to align regions from their analyses. In addition Li and Gouy (1989a) used elements of inferred secondary structure of rRNA to guide their alignments.

The base composition of these sequences is like those of Gouy and Li (1988) which are described extensively in section 3.6.1.

1.9.2 A long stretch of mtDNA from apes

The second data set is relevant to resolving the relationships and divergence times of the African Hominoids, human (*Homo sapiens*), chimpanzees (*Pan troglodytes* or the common chimp and *Pan paniscus*, the pygmy chimp) and gorilla (*Gorilla gorilla*). For this we chose the nearly 5 kilobase mtDNA sequences of Horai *et al.* (1992) comprising the aforementioned species plus orangutan (*Pongo pygmaeus*) and siamang (*Hylobates syndactylus*). This mtDNA is expected to show a very high degree of variability of rates across sites as coding regions (1st, 2nd and 3rd positions), tRNA genes, plus short non-coding regions are all included. Indeed some authors suggest that the 3rd position sites are so saturated with change over the period spanned by this tree (last common ancestor of these taxa was approximately 16 million ago) that they are unreliable and were therefore excluded from an earlier analysis (Horai *et al.* 1992). Here we include all sites (except the less than 40 sites showing deletions) to test how well our model can deal with such an extreme situation. In addition, mtDNA shows a high and often difficult to estimate, transition to transversion ratio, which combined with distinctly unequal nucleotide composition could potentially mislead phylogenetic methods. We consider how well the extensions to Hadamard conjugation models developed in this thesis deal with this situation. Initially we will be considering a four taxon subset (human, chimp, gorilla, orangutan), but later in chapters 5 and 6 use the whole data set in the context of branch and bounding ML and estimating divergence dates.

Table 1.1 Base composition and singleton changes in Hominoid mtDNA data

	Base composition				Singleton changes					
	const.	prop.	var.	prop.	H	C	P	G	O	S
A	1129	0.323	353	0.251	4	3	1	4	22	31
C	969	0.277	543	0.385	11	4	6	12	68	68
G	522	0.149	126	0.089	31	11	6	38	60	46
T	880	0.251	387	0.275	45	14	11	58	78	104
totals	3500	1	1409	1	91	32	24	112	228	249

Note that throughout this thesis when discussing these sequences the following abbreviations are often used: H for human, C for common chimp, P for pygmy chimp, G for gorilla, O for orangutan, and S for siamang.

Table 1.1 gives details of the six sequences. Base composition between species is near stationary, but is marked by a low content of G. There is a marked difference in the frequency of G and C between the constant sites (all taxa the same state) and those sites showing variation. Later in section 3.6.1. we give a X^2 Pearson type test for the significance of this difference, and in this case it is highly significant ($X^2 = 84.675$, d.f. = 6, $p = 1$ in 10^{16}). The base composition of the variable sites appears more uneven than that of the constant sites, and this can only partially be explained by their smaller number (and thus larger stochastic fluctuation). The singleton

changes are reasonably even across taxa, and are marked by a rarity of changes to state A, and a predominance of changes to state T. This is somewhat surprising as A is one of the most frequent states, while T is one of the rarer states. This probably due to a high rate of change in and out of T, particularly via C → T and T → C transitions which are marked in mammalian mtDNA (e.g. Horai *et al.* 1992). The frequency of these singleton changes is used later in chapter 5 when looking at branch and bound of maximum likelihood. (Note, the frequency of singleton changes is how often each taxon shows a different state to all others taxa, irrespective of their states, and is not generally equal to the number of singleton, or sites showing variation but uninformative to unweighted parsimony).

Nuclear encoded psuedogene sequences are used in section 5.4 and are described in more detail there. They are the same hominoid φη sequences used by Hendy *et al.* (1994), which are similar to the set used in Miyamoto *et al.* (1987). A detailed overview of these and additional β-globin region sequences is given in Bailey *et al.* (1992).

1.9.3 Gouy and Li's alignment of diverse 16S-like rRNA sequences

A third data set is analysed primarily in chapter 3. It is the set of aligned 16S-like rRNA sequences used by Gouy and Li (1989a) to evaluate evidence for the monophyly of the archaebacteria. This data was further edited to remove any site with a deletion amongst any of the 28 taxa included, leaving exactly 800 sites. A full description of the taxa from which the sequences came is given in the caption to figure 3.12. A full description of the base composition of these sequences is given in section 3.6.1. The results from this data set serve as an interesting contrast to the analyses of a very similar set of 16S-like sequences analysed by Lake (1988). In addition the aligned sequences used by Gouy and Li (1988) include the very early amitochondrial eukaryotes *Giardia* (a diplomonad) and *Vairimorpha* (a microsporidian). This allows us to evaluate conflicting evidence from Vossbrinck and Woese (1986) and Sogin *et al.* (1989) regarding which of these groups is more anciently diverging.

1.10 OVERVIEW AND COMPUTER SOFTWARE USED

Computer software: A variety of computer software was used. Basic functions such as calculating base frequencies used the programs "Prepare" and "Trees" (Penny *et al.* 1993) which runs under DOS, and "MacClade 3.0" (Maddison and Maddison 1993) which runs on the Apple Macintosh. For Hadamard conjugations, the program "Prepare" (Penny *et al.* 1993) was used to convert the aligned sequences into **s** vectors. Usually these vectors were then read into programs I wrote myself using the features of the spread sheet "Excel" (versions 2.0 and 5.0, by Microsoft). In chapters 4, 5, and 6 simulations using Hadamard conjugations were performed after assisting D. Penny to make mostly minor modifications to "Hadtree" (Penny *et al.* 1993). Statistical analysis of such simulations in chapter 4, involved use of the program Splus running on Sun Unix workstations. General tree estimation used various programs in the PHYLIP 3.5 package (Felsenstein 1993) running under DOS, the program PAUP 3.1 (Swofford 1993), and

most recently test versions of PAUP* (Swofford 1995), both running on the Macintosh. All other programs were written by myself, using primarily the features of Excel.

Overview: This thesis has a number of themes running through it. One of the main themes is developing Hadamard conjugations, and relating them to other methods of analysis. After the general outline in chapter 1, the Hadamard conjugation is extended in chapter 2 to accommodate different evolutionary rates across sites, and considers some distributions which might approximate these. A new set of order 2^{t-1} Hadamard conjugations for use with 4-state data are presented in the appendices to this chapter. Chapter 3 considers applications of allowing for unequal rates across sites in distances, especially the LogDet transformation, which is also studied closely. David Swofford has recently incorporated some of these findings in the computer program PAUP*. The first part of chapter 4 (up to section 4.4) evaluates the variances and sampling properties of the i.r. γ vector and is published in Waddell *et al.* (1994). The following part of this chapter evaluates what happens when there are unequal rates across sites. Various ways of reducing the variance of Hadamard conjugation corrected data are also considered. The last part of this chapter deals with similar statistical issues for the related, but quite distinct, distance Hadamard. This work could be considered three chapters under one heading.

The first part of chapter 5 deals with criteria for choosing trees. This section also uses the intermediate properties of the Hadamard conjugation to compare and contrast usually distinct methods such as weighted least squares on distances, and maximum likelihood from sequences. The next part of this chapter deals extensively with maximum likelihood models when rates across sites vary, and when the data can be considered to be the sum of a number of different processes. One important result is maximum likelihood phylogeny estimation for regions that have undergone recombination. Following this is a large section on how all methods become inconsistent with unequal rates across sites (part of this work is used to illustrate Lockhart *et al.* 1996). The next section looks at the opposite problem, that of overcorrecting the data, and is appropriately called the "anti-Felsenstein zone," as it was Felsenstein (1978) who first pointed out the problems that undercorrection can cause. Another important finding, developed especially in the appendix to chapter 5, is the issue of more than one tree giving the same sequences when rates vary across sites. Chapter 6 focuses upon statistical tests, describing new tests, and making comments upon tests already being used. Chapter 7 is a brief summary chapter.

To follow the theme of the Hadamard conjugation, you would principally read the relevant parts of the introduction chapter, then chapter 2, the first part of chapter 4, the first part of chapter 5, and the first part of chapter 6. To follow the main theme of maximum likelihood, you would read the relevant parts of the introduction, the basis for chapters 2 and 4, then read the relevant sections in chapter 5, and parts of chapter 6. To follow the theme of LogDet you would read the introduction and chapter 3.

CHAPTER 2:

EXTENDING HADAMARD CONJUGATIONS TO MODEL UNEQUAL RATES ACROSS SITES

2.1 INTRODUCTION

An important property of any statistical estimator is that it be *consistent*, that is, as more data are added the estimator converges to its true value (chapter 1). In the case of tree building methods, Felsenstein (1978) defined consistency as convergence to the true unweighted tree (the tree that generated the data) as longer sequences are gathered. As discussed in section 1.6, consistency may also be applied to recovering the true weighted tree, i.e. recovering rates of evolution exactly. A major obstacle to a method recovering the true weighted, or unweighted, tree from biological data is the occurrence of parallel or convergent changes during evolution. In the case of DNA sequences, this is manifest as multiple substitutions at a site (or multiple hits).

In order to estimate the number of parallel or convergent changes a model of character state change (a *mechanism*) is required. In the case of DNA sequences this may be described by a transition matrix of substitution rates. An early example of the application of such a model, was the Jukes-Cantor distance correction (Jukes and Cantor 1969, formalised by Kimura and Ohta 1972). In this model the implied total number of character state changes between two sequences was estimated from the observed proportion of substitutions according to a simple Poisson model of nucleotide substitution. If the model used matches the process of nucleotide substitution then these estimated distances will be tree-additive as the sequence length goes to infinity. There are tree building methods which are always consistent when supplied with tree-additive distances (Felsenstein 1988 and references therein); for example, methods that choose the tree which minimises a sum of squares between observed and predicted distances. It is important to recall from chapter 1 that it is required for a distance to be additive upon all possible trees, since in chapter 5 we show that distances can be "locally additive" upon certain incorrect trees.

While distance corrections provide one way of dealing with multiple substitutions at a site in a sequence, they have limitations. One of these is the loss of information in going from sequence patterns to a matrix of pairwise distances (Penny 1982). For example, the number of distances between sequences grows as order t^2 (where t is the number of sequences being compared) whereas the number of sequence patterns grows exponentially (Penny 1982, Steel *et al.* 1988). We expect that this mapping down from sequences to distances will, in some circumstances, lose information relevant to reconstructing the true phylogeny (we find further examples of this in chapter 5). In addition, the relationships implied in a distance matrix can be difficult to describe or visualise, whereas systematists are used to working with and diagnosing character state patterns when classifying organisms. Consequently it is most desirable to have

methods that can correct the frequencies of sequence patterns for multiple changes, analogous to the inferred corrections made to pairwise distances prior to tree selection.

Recently Hendy and Penny (1993) have shown that it is possible to correct sequence patterns for inferred multiple substitutions by using a series of easily computable transformations (as discussed in chapter 1). These Hadamard conjugation methods have recently been extended to four or more character state models by Steel *et al.* (1992), Szekely *et al.* (1993), and Hendy *et al.* (1994). While Hadamard conjugations are consistent under i.r. (identical rate) and i.i.d. (all sites evolving identically and independently) models, it is known that in DNA sequences there are often large differences in substitution rates at different sites. The term consistent applied to Hadamard conjugations means that if the sequences evolved according to the mechanism specified, then as sequence length goes to infinity, the output of the Hadamard conjugation converges to a vector description of the tree that generated the data.

In coding regions, variation of rates between sites is often pronounced between first, second and third position sites, partly due to different degrees of amino acid code degeneracy allowing different proportions of neutral substitutions. Further, even within either first, second and possibly third positions, there will often be large differences in the probability of change due to differential stabilising selection upon the amino acid positions coded for (e.g. Nei 1987). Fitch and Markowitz (1970) go even further and argue that many sites in coding regions are completely invariant, i.e. they cannot change at all within the group of organisms being considered because such changes are lethal. It is now known that such unevenness of rates can also be present in non-coding regions, including psuedogenes, due to unequal mutation rates. For example CG dinucleotide sites appear to be over 100 times more likely to mutate than the average site in primate nuclear DNA (Perrin-Pecontal *et al.* 1992).. Variations of rates across sites may result in serious underestimates of pairwise distances when an i.i.d. model correction is used (Golding 1983, Olsen 1987, Nei and Jin 1990). This underestimation increases as the observed proportion of substitutions increases. We show later in this chapter, that the degree of underestimation can be greater than the difference between the observed and the inferred number when applying the same i.i.d. correction to commonly used sequence data.

Main themes of this chapter are to:

1) Address the general problem of how to derive path length correction formulae when rates vary across sites.

2) Describe models to approximate a real sequences distribution of rates across sites (especially using well known unimodal statistical distributions).

3) Consider especially two proposals regarding the form that the distributions of rates across sites may take in coding regions, these being:

(A) rates across sites are described by a gamma distribution (Uzzel and Corbin 1971).

(B) there is a proportion of sites which cannot change for functional reasons, the invariant sites (Fitch and Markowitz 1970).

4) Modify Hadamard conjugations to take such distributions into account by deriving the appropriate pathset correction formulae and prove that such modifications are exact under specific models. These conjugations can either "correct" the observed data, or estimate the probability of data when specifying the models parameters.

We illustrate the importance of these extensions to the recovery of evolutionary trees, and the estimation of edge lengths with real and model-derived data.

Note that these extensions to the Hadamard conjugation where originally developed and implement by the author, then in cooperation with Dr Mike Steel a mathematical proof was developed (see Steel *et al.* 1993c, and appendix 2.2). At the same time these Hadamard conjugations were also extensively used to implement maximum likelihood tree selection with a distribution of rates across sites (Waddell and Penny 1995, accepted in 1993, and see also chapter 5).

2.2 HADAMARD CONJUGATIONS REVIEWED

This section describes in detail Hadamard conjugations, and how they are calculated. The following section 2.3 then explains how Hadamard conjugations can be extended to models with a distribution of rates across sites, which constitutes the start of novel work in this chapter.

A new approach to transforming data, prior to selecting phylogenetic trees, is spectral analysis (Hendy and Penny 1993) where aligned sequences are corrected for implied multiple substitutions (hits) without reduction to just the pairwise distances. We will first describe the method using a simple 2 character state model, while later in the chapter we will introduce then expand on the 4-state Hadamard conjugation of Hendy *et al.* (1994). When using a Hadamard conjugation the proportions of sites showing each distinct sequence pattern are expressed as a vector **s** (**s** from sequences). When dealing with 2-state characters, **s** is also called a vector of bipartitions, or splits, as the character states define two exclusive subsets. The Hadamard conjugation is made up of three steps: vector **s** is multiplied by a Hadamard matrix (a discrete Fourier transform) to give a vector of observed "pathset lengths" (generalised distances), which are corrected for multiple hits, and finally the inverse Hadamard transform converts them back to reweighted sequence patterns in the vector γ. With 2-state data these generalised distance corrections are made under the Cavender-Farris model (Cavender 1978, Farris 1973), while with 4-state data they are made under a general form of Kimura's 3ST model (Kimura 1981). Each entry in γ not associated with an edge on the tree that generated the data is an invariant (in the sense of Cavender 1978) with an expected value of 0 as the number of sites, c, tends to infinity, i.e. $c \rightarrow \infty$ (Hendy and Penny 1993, Steel *et al.* 1993c). Each entry in γ associated with an edge in the tree takes on a positive value equal to the expected number of substitutions per site along that edge (except for γ_0, which is minus the sum of these values).

From the adjusted weighted bipartitions in γ it is possible to choose a set of up to 2t-3 compatible bipartitions, that is a tree, by optimality criteria such as maximum parsimony,

compatibility, or closest tree (Hendy 1991). These criteria, which can be inconsistent when applied to uncorrected data, are always consistent when applied to γ (Steel *et al.* 1993b), as long as γ is consistent. A useful feature of a Hadamard conjugation is that it is invertible; that is given a weighted tree we can use the conjugation to calculate the probabilities of observing each bipartition pattern in sequences (the vector **s**(T)) generated under the appropriate models. We can then use these vectors of sequence pattern probabilities as the starting point for simulation studies (e.g. Charleston and Hendy 1993, Waddell *et al.* 1994). With real data they can be used to check the goodness of fit between data and model, or to perform calculations of the likelihood of the data under a specific model, as described in later chapters. Perhaps one of the Hadamard conjugation's greatest advantages is that it has a relatively simple structure which facilitates the application of statistical methods.

2.2.1 Definitions and a worked example

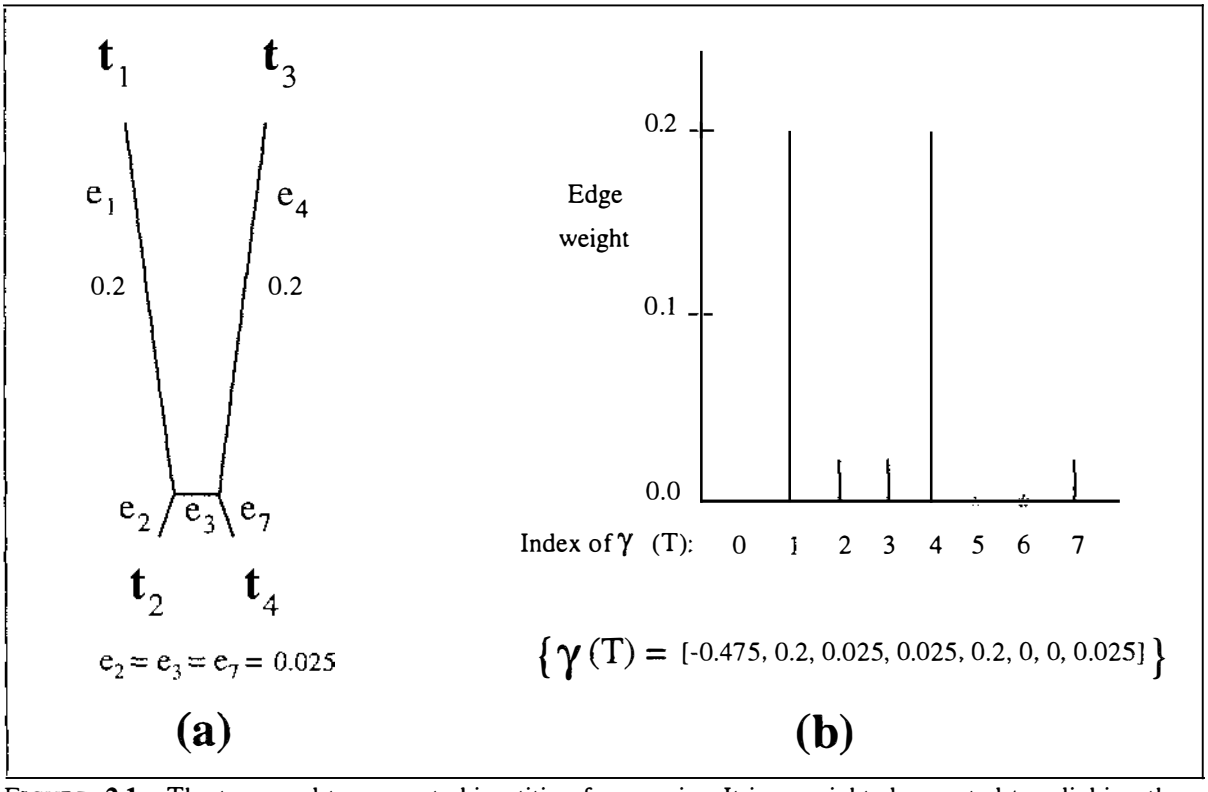


FIGURE **2.1a**. The tree used to generate bipartition frequencies. It is a weighted unrooted tree linking the four taxa, t_1 to t_4 , and is drawn to scale. The edges (e) are labeled according to the binary indexing of Hendy and Penny (1993). This tree, like the example in Felsenstein (1978, fig. 1), has only two different weights on the edges. Edge lengths are expressed in expected number of changes per site. **2.1b** The tree re-expressed in vector form as γ(T) (lower), and in graphical form as a weighted spectrum (upper) using the same indexing e.g. $e_1 = \gamma(T)_1$ etc. The entry for $\gamma(T)_0$ (which is minus the sum of all other γ_i) is omitted from the spectra for ease of presentation. Entries γ_5 and γ_6 have weight 0 because corresponding edges do not occur in this tree (they index the internal edges relating to the two other resolved unrooted four taxon trees). Applying the Hadamard conjugation to γ(T) gives **s**(T) a vector of the probabilities of the character bipartitions (see equation (1)). Vector **s**(T) is shown in last column of table 2.1.

Table 2.1 Example of the calculation of s(T) using the tree derived γ vector in figure 2.1

Index[a]	Pattern[b]	γ(T)	**H** [c]								ρ = Hγ(T)	**r** = exp(ρ)	**s**(T) = H^{-1}**r**
0	0000	-.475	1	1	1	1	1	1	1	1	.00	1.000	.6479
1	0001	.200	1	-1	1	-1	1	-1	1	-1	- .50	.6065	.1283
2	0010	.025	1	1	-1	-1	1	1	-1	-1	- .15	.8607	.0200
3	0011	.025	1	-1	-1	1	1	-1	-1	1	- .45	.6376	.0226
4	0100	.200	1	1	1	1	-1	-1	-1	-1	- .45	.6376	.1283
5	0101	.000	1	-1	1	-1	-1	1	-1	1	- .85	.4274	.0258
6	0110	.000	1	1	-1	-1	-1	-1	1	1	- .50	.6065	.0070
7	0111	.025	1	-1	-1	1	-1	1	1	-1	- .90	.4066	.0200

This table shows intermediate steps in calculating the observed sequence pattern probabilities **s**(T) from γ(T).

[a] Assign each taxon an integer from 1, 2,..., t (here t= 4). The pattern of two coloured states at a site is expressed as an ordered string (the lowest numbered taxon at the far right) with a 1 indicating which taxa have a different state to the last t-th taxon. These strings are also binary numbers which have Arabic equivalents. For example the pattern 0101 is the binary equivalent of 5. To do the reverse and derive a bipartition pattern from its Arabic index, simply express it as a binary number, filling out any extra places to the left with 0's until there are t digits.

[b] Notice that there are only 8 and not 16 distinct patterns since the Cavender-Farris model implies that the frequency of sites with pattern 0110, for example, is equal to the frequency of sites with pattern 1001 and so on, making their separate consideration unnecessary.

[c] To visualise what the Hadamard matrix is doing the reader may wish to rewrite the matrix replacing 1's with 0's then -1s with 1 to give matrix **Y**. Then confirm that r_j is $1-2d_j$, where $\mathbf{d} = \mathbf{sY}$. For example r_1 counts proportion of sites (1 same as 4) minus prop. ($1 \neq 4$), so $r_1 = p(1 = 4) - p(1 \neq 4) = (1 - p(1 \neq 4)) - p(1 \neq 4) = 1-2p(1 \neq 4)$.

Following Hendy and Penny (1993) let our weighted tree, T of t taxa, be described by the vector of weighted edge lengths γ(T). The tree T we use to illustrate our methods (fig. 2.1a) is similar to Felsenstein's (1978) example with only two different edge weights chosen so that parsimony applied to uncorrected sequence data will converge to the wrong tree. The edge lengths (weights) are the total expected number of changes per nucleotide site on that edge (counting multiplicity of changes). These are represented by γ(T) of figure 2.1b. The mechanism of character state change used here is that of Cavender (1978) and Hendy and Penny (1989 and 1993), the 2-state analogue of the Jukes-Cantor equation. The two character states could for example refer to the two chemical classes (purines and pyrimidines) of nucleotides in DNA. Substitutions at sites occur independently of each other and the probability of substitution is identical at all sites, so the mechanism of change is i.i.d. (independent and identically distributed). Also changes are independent across edges in the tree.

Following Hendy and Penny (1993) we apply the Hadamard conjugation to γ(T) to obtain **s**(T) the vector of the probabilities for each of the possible 2^{t-1} site patterns (bipartitions), that is

$$\mathbf{s}(T) = \mathbf{H}^{-1}(\exp(\mathbf{H}\gamma(T))). \qquad (2.2.1\text{-}1)$$

In this formula **H** is a symmetric Hadamard matrix, $\mathbf{H}^{-1} = 2^{-(t-1)}\mathbf{H}$, and the exponent function is applied component-wise. We define the Hadamard transform as the multiplication of a vector by **H**. A Hadamard conjugation involves a Hadamard transform, then an operation(s) on the resulting vector, followed by the inverse Hadamard transform. Table 2.1 gives a numerical example of a Hadamard conjugation, and shows the form of H for four taxa. We define **H** as the $(t-1)^{th}$ Kronecker power of the matrix, $\begin{bmatrix} 1 & 1 \\ 1 & -1 \end{bmatrix}$. The simple recurring structure in this matrix can be exploited to obtain an exponential order increase in the computation speed of the Hadamard transform by the method known as the Fast Hadamard Transform (Tolimieri *et al.* 1989, Hendy and Penny 1993). This decreases the number of operations in calculating a Hadamard transform from order $(2^{t-1})^2$ to order $(t-1)(2^{t-1})$. The Hadamard conjugation is fully invertible (meaning that it losses no information) and so can be rewritten as,

$$\gamma(T) = \mathbf{H}^{-1}(\ln(\mathbf{H}\mathbf{s}(T))) \qquad (2.2.1\text{-}2).$$

For convenience and interpretability the conjugation can broken into three steps (Hendy and Penny 1993),

$$\mathbf{r} = \mathbf{H}\mathbf{s}; \qquad (2.2.1\text{-}3)$$

$$\rho = \ln \mathbf{r}; \qquad (2.2.1\text{-}4)$$

$$\gamma = \mathbf{H}^{-1}\rho. \qquad (2.2.1\text{-}5)$$

Each component of **r** is equal to,

1 - {twice the proportion of sites where the number of character states assigned to the taxa in that pathset is odd} (see table 2.1).

As such an observed 'length' for a pathset with four end points would be,

1 - {twice the proportion of sites which had one of the four taxa's states different to all others},

while for a six taxa pathset it would be

1 - {twice the proportion of sites which had 1 or 3 states different to all others in that subset of six taxa}.

With the i.r. model, the *i*-th component in ρ, ρ_i, is just the natural logarithm of r_i. A subset of pathsets are single paths (pairwise distances), so under a symmetric 2-state Poisson mechanism of character state change the inferred distance is -1/2ln(1- 2 × observed distance) or $-1/2\ln(r_i)$ where *i* indexes a pathset with just two end points (Hendy *et al.* 1987). The -1/2 coefficient does not appear in equation 2.1, because it cancels with a -2 coefficient from the $\mathbf{H}^{-1}$ transform to give $\gamma(t)$ in the correct units. The aforementioned distance correction equation is the two character state analogue of the four character state equation commonly known as the Jukes-Cantor, which is the solution of a continuous time first order Markov process (e.g. see Gojobori *et al.* 1990).

Following Hendy and Penny (1993), to index a bipartition pattern first assign each taxon a successive integer (i) from 1, 2, ... , t, and the index of all taxa except the last one is 2^{i-1}. The indexing of any pattern is just the sum of the indices of all the taxa that differ in state from the t-th taxon, for example with four taxa the pattern 0 1 1 0 (which could be RYYR, where R is a purine and Y is a pyrimidine when reading down a column of aligned data) has the index $2^{2-1} + 2^{3-1} = 2^1 + 2^2 = 6$ (see table 2.1 for further examples). With 5 taxa the pattern 0 1 0 1 1 has index $2^0 + 2^1 + 2^3 = 11$. The pattern 1 0 1 0 0 has the same index because character states are indexed relative to the state of the last taxon, t.

The "Fast Hadamard" form of the Hadamard transform is the fastest way known to invertibly (hence without losing any information) turn bipartitions into a set of pathset lengths without needing to specify a particular tree or graph. Single paths (i.e. pairwise distances) as we have already mentioned can be interconverted between observed and "true" lengths (counting multiplicity of changes), and under certain mechanisms of substitution, so too can non-intersecting sets of paths be interconverted between end point labelings (which we will refer to as observed pathset lengths, or observed generalised distances) and true lengths. So the Hadamard conjugation is converting bipartition data into pathset lengths, which are next corrected for implied multiple changes in the data, followed by applying the inverse Hadamard transform to calculate 'corrected' character state patterns. Note, it can often help to think of pathset lengths as generalised distances, which can be either additive on a tree (in ρ), or an observed quantity which hides multiple substitutions at a site (in $\mathbf{r}$).

Presently the i.r. Hadamard conjugation is known to be consistent for the Cavender / Farris 2-state model, recovering exact rates on edges in the generating tree, with all other entries going to zero (Hendy and Penny 1993). With two or more states, the Hadamard conjugation is exact only if the set of changes on the nucleotides forms a Boolean (and hence also an Abelian) group (Székely *et al.* 1993). A Boolean group implies that all members of the group (the 4 states) can undergo the same types of change, and application of the same type of change twice in a row brings back the identity, or no apparent change. This condition also implies that all transition matrices on the tree are symmetric (this rules out the 2-state asymmetric model which Hendy and Penny 1993 had mistakenly assumed would work). With 4-state characters, the most general model obeying this requirement is the generalised Kimura 3ST model (Kimura 1981, Evans and Speed 1993) where the rates of transitions and the two types of transversions can vary independently on all edges of the tree (Steel *et al.* 1992). All these models have been defined under i.i.d. conditions with identical rates at all sites, with the additional stipulation that each site evolves independently between edges of a graph (substitutions on one edge do not alter the probability of a substitution on another edge). These conditions imply that the evolution of sites must follow a tree.

While correlation of changes amongst characters is a concern to all phylogeneticists, Hadamard conjugations (and the other main classes of phylogenetic methods) can still be consistent under such circumstances. Consistency is assured as long for some reordering of the sites, the correlation of substitution probabilities between sites i and j decreases at a sufficient

rate, because then in the limit as the sequences become infinitely long $\hat{\mathbf{s}} \rightarrow \mathbf{s}(T)$ (e.g. see Bernstein's theorem in Rényi 1970 p.379)(where $\hat{\mathbf{s}}$ is just a vector of the observed proportions of sites with each distinct site pattern). As a specific example of the conditions required, generate an order of the sequence sites (0 up to infinity) that reflects decreasing correlation of site substitutions (not necessarily the direct sequence order, but preferably one that best takes into account secondary and tertiary interactions of molecules). As we take pairs of sites further apart in this new ordering, then as long as the correlation between all pairs of sites x and y eventually becomes less than $\frac{1}{|x-y|}$ (where x - y is just how many sites apart they are) we get convergence of the observed site pattern probabilities to the exact values expect in $\mathbf{s}(T)$ under the i.i.d. model. One way of looking at this convergence is that nonindependence is only effectively occurring between a finite number of sites, which is overcome as our sequences become longer (Rényi 1970, describes various limits of the rate at which correlations must decrease in order to make this statement true). The consistency of phylogenetic methods given correlated changes between sites has been noted before (e.g. Hasegawa *et al.* 1985, Steel *et al.* 1993c). While this convergence result is reassuring, as Hasegawa *et al.* (1985) mention the most serious effect of correlation is not upon the consistency of a method, but rather upon associated statistics for finite sequence length. In chapter 4, we suggest a possible solution to this problem in the case of Hadamard conjugations.

2.3 HADAMARD CONJUGATIONS WITH UNEQUAL RATES ACROSS SITES

It is very desirable to extend Hadamard conjugations to allow for a distribution of rates across sites, since it is known that this is an important factor in biological data due to different degrees of selection on sites in a DNA sequence (Nei 1987). It has already been shown that any variability of rates across sites will result in an underestimation of the true distance from the observed sequence mismatches if this factor is not taken into account (Golding 1983). This underestimation occurs in a non-linear way (see figure 2.2), resulting in non-additive distances which in turn may cause distance based tree building methods to be inconsistent (Felsenstein 1988).

2.3.1 Discrete distributions

If the distribution of rates at different sites is discrete (as shown in figure 2.2), then an exact measure of the evolutionary distance (δ) is a sum of the corrected distances for each distinct rate, multiplied by its proportion. That is, a weighted sum of the expected distance for each rate class, with the weights being the probability of a site belonging to a particular class. That is,

$$\delta = \mathrm{E}[P(\mathrm{d}\lambda_i)] = \sum_{i=1}^{n} P(\mathrm{d}\lambda_i) p\lambda_i \qquad (2.3.1\text{–}1)$$

where $P(d_{\lambda i})$ is a consistent distance correction applied to the observed proportion of changes (d) at sites with rate λ_i, $p\lambda_i$ is the proportion of sites with rate λ_i and n is the number of different classes of rates. Notice then, that if we are to infer consistently the total number of substitutions between two sequences, then in addition to correcting for the substitution process in each class of sites we need to know both their relative rates and their relative proportions. Since we are constrained by the fact that the proportions of each class must sum to 1, and that the mean rate is 1 (i.e. $p\lambda_i$ also sums to 1), then for n classes of sites we need $2n$ -2 parameters to describe the distribution.

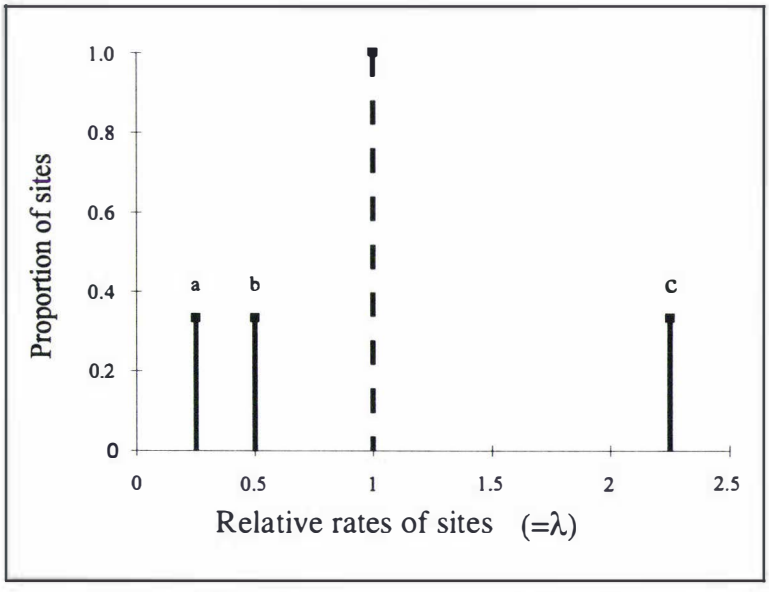


FIGURE 2.2 An example of different sites in a sequence having different rates (labeled a,b,c) which give an average rate of exactly 1 substitution per site. The sites with rates a, b, c could, for example, correspond to second, first and third positions in a protein coding sequence (having relative rates of 1:2:9). Here the proportion of sites in each class is equal at one third, but one could model other relative proportions, given the sole constraint that they sum to one. The dashed line marks the most commonly employed distribution of rates across sites, the identical rate at all sites. Note the average rate has been normalised to 1.

2.3.2 Continuous distributions

If the distribution of rates across sites is continuous, then like the discrete case of equation (2.3.1-1) an additive measure of the true distance between sequences when there are different rates at different sites is the integral

$$\delta = \mathrm{E}[P(\lambda)] = \int_{\lambda=0}^{\infty} P(\lambda) f(\lambda) \mathrm{d}\lambda \qquad (2.3.2\text{–}1)$$

(Golding 1983, Jin and Nei 1990). Here $P(\lambda)$ is the distance correction formula, while $f(\lambda)$ is the probability density function (p.d.f.) of the distribution of rates across sites. Note that for this form of estimator to be consistent across a path or pathset, the intrinsic rate at a site, relative to the rates at other sites, must be fixed. In the case of the Cavender (1978) symmetric 2-state identical rates model $P(\lambda)$ (the corrected distance) is equal to $-1/2 \ln (1-2d_{fg})$, where d_{fg} is the observed proportion of substitutions between sequences f and g and ln is the natural logarithm. So we can extend the 2-state Hadamard conjugation to allow unequal rates across sites, such that

$$E[\rho_i] = \int_{\lambda=0}^{\infty} \ln(r_i(\lambda)) f(\lambda) d\lambda \; = M^{-1}(r_i). \qquad (2.3.2\text{-}2)$$

For reasons discussed in the next section we label $E[\rho_i]$ the function $M_x^{-1}(r_i)$ (the inverse of the moment generating function of variable X). This estimator is consistent for a set of models where the sites in the sequence have only one underlying distribution of rates across sites, $f(\lambda)$; the intrinsic rate at a site is randomly chosen from this distribution; the intrinsic relative rates of sites do not change over the tree (i.e. the relative rates of all sites remains constant), but the rate at all sites may be proportionately increased or decreased on any edge in the tree. (In appendix 2.1 we show, via a counter example, why these conditions must hold in order to guarantee consistency). We will define this type of model as still being i.i.d.; we have simply relaxed the additional requirement that all sites have an identical rate (an i.r. model).

Under the mechanisms of character state change for which Hadamard conjugations are consistent we can apply the same path set transformation to all elements in **r**, or its inverse to all elements in ρ. Allowing for variation of rates across sites we rewrite the Hadamard conjugation as either

$$\mathbf{s}(T) = \mathbf{H}^{-1}(M_x(\mathbf{H}\gamma(T))) \qquad (2.3.2\text{-}3)$$

or

$$\gamma(T) = \mathbf{H}^{-1}(M_x^{-1}(\mathbf{H}\mathbf{s}(T))) \qquad (2.3.2\text{-}4)$$

depending on whether we wish to calculate the probabilities of sequence site patterns (**s**) or alternatively correct sequence data for multiple hits (γ).

We now describe why this transformation is consistent for any combination of parameters meeting the assumptions. Firstly we have that the Hadamard transform $\mathbf{r} = \mathbf{Hs}$ is an orthogonal linear transformation (Hendy and Penny 1993). It follows then that **r** is additive in the sense that $\mathbf{r}_{\lambda(1+2)} = p_1\mathbf{r}_{\lambda 1} + p_2\mathbf{r}_{\lambda 2} = p_1\mathbf{H}\mathbf{s}_{\lambda 1} + p_2\mathbf{H}\mathbf{s}_{\lambda 2}, = \mathbf{H}(p_1\mathbf{s}_{\lambda 1} + p_2\mathbf{s}_{\lambda 2})$ where $\lambda 1$ and $\lambda 2$ the rates of change for two classes of sites which have otherwise evolved by the same mechanism according to the same weighted tree; p_1 and p_2 are the proportion of sites that have rate $\lambda 1$ and $\lambda 2$ respectively. By induction it follows that this additivity holds for any sum of sites evolving by the same model but for having different absolute rates. Integrals (such as 2.3.2-2) will always be defined (as long as a pathset length is not infinite) since the probability density function (p.d.f.) of rates is continuous with a finite mean (here specifically set to one), consequently the p.d.f. must ultimately decrease faster than the function 1/x (which lies on the boundary of finite integrals that go to + ∞). The reason we have the mean of the p.d.f. set to one is so that we recover the pathset lengths as the expected number of changes per site and not some multiple.

We now check that the above integral (2.3.2-2) describes a set of unique and invertible transformations for every possible distribution of λ. When **s**(T) is defined as a vector of pattern probabilities, it follows that under the 2-state Poisson model the maximum observed divergence will be 0.5, and $0 \leq r_i \leq 1$ (Hendy and Penny 1993), and $\ln(\mathbf{r}) = \rho$ is monotonically increasing in this range ($r \neq 0$). As ρ_λ may be viewed as the sum $\rho_{\lambda 1} + \rho_{\lambda 2} + \rho_{\lambda 3} \ldots$, then since ρ_λ is a sum of

monotonic functions it too must be monotonic. Monotonicity in turn guarantees that for each x there is a unique y, so integral (2.3.2-1) is always invertible i.e. we can recover $\mathbf{r}_\lambda$ from ρ_λ given the distribution of λ (its p.d.f.) and vice versa. It is also true that for every different distribution of the λ there will be a unique transformation from **r** to ρ since each of the terms in the sum $\ln(\mathbf{r}_{\lambda 1}) + \ln(\mathbf{r}_{\lambda 2}) + \ln(\mathbf{r}_{\lambda 3})$ ... is a unique and finite monotonic function on the defined range of r_i. Thus the functions defined by the integral 2.3.2-2 are always continuous and analytic (i.e. have a derivative at each point) for the domain of r_i corresponding to finite pathset lengths ($0 < r_i \leq 1$). Lastly just as $\rho_\lambda = \mathbf{H}\mathbf{s}_\lambda$, then it holds that $\gamma_\lambda = \mathbf{H}^{-1}\rho_\lambda$ is also exact since $\mathbf{H}^{-1}$ like **H** is an orthogonal and invertible transformation.

This concludes our description of a proof that using formula 2.3.2-1 to go from **r** to ρ (or vice versa) will give an exact result given that: the assumptions for an i.i.d. and identical rates Hadamard conjugation hold (Hendy and Penny, 1993, Székely *et al.* 1993, Hendy *et al.* 1994). And also that:

(a) The relative rate of site *i* to site *j* is constant during the evolution of the sequences, for each pair *i* and *j*, and

(b) The distribution of rates across sites is known.

In appendix 2.2 we give a mathematical proof of the consistency of Hadamard conjugations under this variable rates across sites model, which is derived with modification from one of our proofs in Steel *et al.* (1993c).

2.3.3 Closed form pathset correction formulae

In correction formula 2.3.1-1 we are required to classify sites into their rate class, while in 2.3.2-1 and 2.3.2-2 we either must be even more precise and classify sites into their exact rate interval, or else use the alternative approach outlined below. In most cases there is only circumstantial knowledge of what the rate at a given site is (often by considering its parsimony reconstruction length on what is hopefully the true tree e.g. Wakeley 1993). A similar approach can be taken with likelihood, for example a recent program by Felsenstein (DNAML 3.5, in Felsenstein 1993) uses an extension of this general approach, classifying sites with a hidden Markov chain model (when we assume that adjacent sites have correlated rates). These approaches can be computationally expensive. The need to a priori classify sites into rate classes can be circumvented if we start with a specified p.d.f. for λ, and predict observed divergence values (i.e go from ρ to **r**). Then it is just a matter of inverting this functional relationship between ρ and **r**, to infer ρ when we know **r**. In some important cases consistent estimators of r_i (given ρ_i and the p.d.f. of site rates) even have closed form inverses which considerably simplify and speed up pathset length transformations. An useful observation is that the transformation from **r** to ρ

$$r_i = \int_{\lambda=0}^{\infty} \exp(\rho_i(\lambda)) f(\lambda) d\lambda \qquad (2.3.3\text{-}1)$$

(where $f(\lambda)$ is the probability density function of λ), can be expressed as $E[e^{Xt}]$, where X is a random variable drawn from $f(\lambda)$, while t is a real valued number (in our case the value of ρ_i). Note that equation (2.3.3-1) can be rewritten as,

$$r_i = \lim_{c\to\infty} \frac{1}{c} \sum_j e^{\rho_i \lambda_j} \qquad (2.3.3\text{-}2)$$

In statistics a function which returns the value of $E[e^{Xt}]$ (where *t* is the argument) is known as the moment generating function, or $M_x(t)$ (Stuart and Ord 1987, Ch. 4), which is effectively what equations 2.3.3-1 and 2.3.3.2 are estimating. This insight facilitates finding closed form expressions for $M_x(t)$ in the literature. For our purpose we equate *t* with the true path length (ρ_i), while X equates with λ_i, a random variable drawn from a p.d.f. (or a probability distribution in the case of discrete rates). An important point is that in our applications the moment generating function applied to any variable with range zero or less (such as ρ_i) is always finite and lies between 1 and 0. This is true even if the moment generating function is not defined for positive values (e.g. as in the case of the lognormal moment generating function).

For estimating edge lengths on trees we require the inverse of equation 2.3.3-1 and 2.3.2-2, to give values that are the expected number of substitutions per sites given observed pathset lengths (or r_i). To recover the number of substitutions per site (and not a multiple of this number) we must ensure that the mean of the assumed distribution of rates across sites is fixed to 1, no matter what shape the distribution may take. In effect we are interested in the distribution of relative rates across sites. In appendix 2.3 we give examples of how to set the mean to 1, and then as an example derive the moment generating function of the uniform distribution. The moment generating function of a distribution is closely related to its "characteristic function" (often abbreviated c.f., see Stuart and Ord 1987, p. 119), while appendix 2.4 details the relationship between these two functions for the purpose of obtaining $M_x(t)$ from c.f.(x)(which is often derived in statistical works). Note again that under the model there is just one fixed distribution of relative rates across sites, which applies to the whole tree.

We now specify one essential, and three desirable, mathematical attributes of continuous distributions of rates between sites:

(1) The p.d.f. has all non-zero values ≥ 0, to correspond to acceptable values of λ (real rates).

(2) It has a closed form expression for the moment generating function.

(3) The moment generating function also has a closed form inverse.

(4) The ratio of the standard deviation to the mean of the random variable is not bounded.

Table 2.2 shows a list of well known statistical distributions that meet the first criteria. Two well known continuous distributions have all four properties. Note that the exponential distribution is a special case of both the gamma and the Weibull distribution when their shape parameter (as listed in table 2.2) is set to 1.

Table 2.2 Common p.d.f.'s, their associated moment generating functions (M) and inverses (M^{-1})

Distribution	par.	p.d.f. (mean set to one)	M	M^{-1}
identical rates-i.r	-	delta function at 1	$\exp(\rho_i)$	$\ln(r_i)$
uniform	b	$1/(2b)$, $1-b \leq x \geq 1+b$ zero elsewhere, $(0 < b \leq 1)$	$r_i = 1/(2b\rho_i)$ $[\exp((1+b)\rho_i)$ $-\exp((1-b)\rho_i)]$	not closed form
gamma (includes the exponential)	k	$(k)\frac{(xk)^{k-1}e^{-(xk)}}{\Gamma(k)}$	$r_i = ((k - \rho_i) / k)^{-k}$	$\rho_i = k\,(1 - r_i^{-1/k})$
inverse Gaussian	d	$(d/2\pi x^3)^{0.5}\exp\{-d(x-1)^2/2x\}$	$r_i =$ $\exp[d\{1-(1-(2\rho_i/d))^{0.5}\}]$	$\rho_i =$ $0.5d[1-\{1-(\ln(r_i)/d)\}^2]$
lognormal	σ^2	$(\mu)\frac{1}{\sigma(x\mu)\sqrt{2\pi}}\exp\left\{\frac{-\ln(x\mu)^2}{2\sigma^2}\right\}$ where $\mu = \exp(0.5\sigma^2)$	not closed form	not closed form
Weibull	w	$(\mu)w(x\mu)^{w-1}\exp(-(x\mu)^w)$, where $\mu = \Gamma(1+1/w)$	not closed form	not closed form
beta (second kind), (includes the F distribution)	p, q	$(\mu)\frac{(x\mu)^{p-1}}{B(p,q)(1+x\mu)^{p+q}}$, where $\mu = \frac{B(p+1,q-1)}{B(p,q)}$, $= p/(q-1)$	(see Phillips 1982)	not closed form

Here the p.d.f. (probability density function) is that of a random variable X which has been standardised, and normalised (mean set to one) (e.g. see Stuart and Ord, 1987, p. 192). The term "par." is the distributions parameter(s), while M is the moment generating function of X, $M_x(t)$, reparameterised to show the relationship of ρ_i to r_i. M^{-1} is the analytic inverse of M. In all cases the shape parameter must be > 0, with the additional restriction for the uniform that it must also be ≤ 1. The gamma and the Weibull give the exponential distribution when their shape parameter = 1. The gamma is the same shape as the Chi-square distribution when *k* is a positive integer (the degrees of freedom in the Chi-square). The shape parameter of the inverse Gaussian has been redesignated d, rather than the commonly used λ to avoid confusion. σ^2 is the variance of the normal distribution (with mean zero) that generates the standardised lognormal distribution. Moment generating functions, and their inverses, are always defined in the ranges of ρ_i and r_i respectively, so if they do not appear here in closed form they can be evaluated numerically.

Because of its shape parameter (*k*) the gamma distribution can mimic a variety of possible biological distributions. As *k* goes to infinity the gamma distribution converges to the i.r. distribution. Figure 2.3 shows how, as *k* ranges form 10 to 0.5, the distribution changes from approximately normal (implying mostly near equal rates) to L shaped with very spread out rates at different sites. A commonly used measure of the spread of a variable is its coefficient of variation (c.v.) which is the ratio of standard deviation / mean (which for distributions in table 2.2 becomes the standard deviation, as the mean is set to one). For the gamma distribution, the c.v. is $1/\sqrt{k} = k^{-0.5}$ (see appendix 2.3). We will use $\Gamma(k)$ to denote a standardised, normalised (i.e. mean set to 1) gamma distribution with shape parameter *k* (and likewise for the other distributions; inverse Gaussian (IG), Beta second kind (F), lognormal (lnN) and uniform (U)).

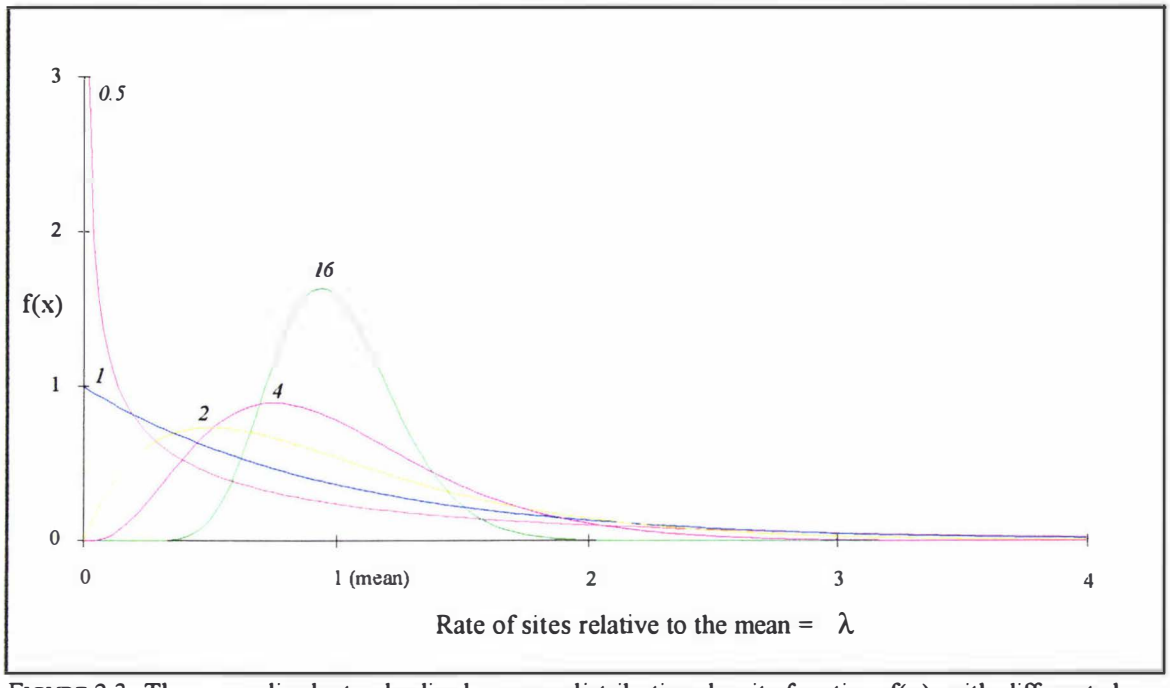


FIGURE 2.3 The normalised, standardised gamma distribution density function, f(x), with different shape parameters (k), as shown associated with the peak of each curve. With k = 1 the gamma distribution equals the exponential distribution. For k less than 1, the gamma distribution may be called "hyper-exponential," to convey its exaggerated shape which is converging to an L shape.

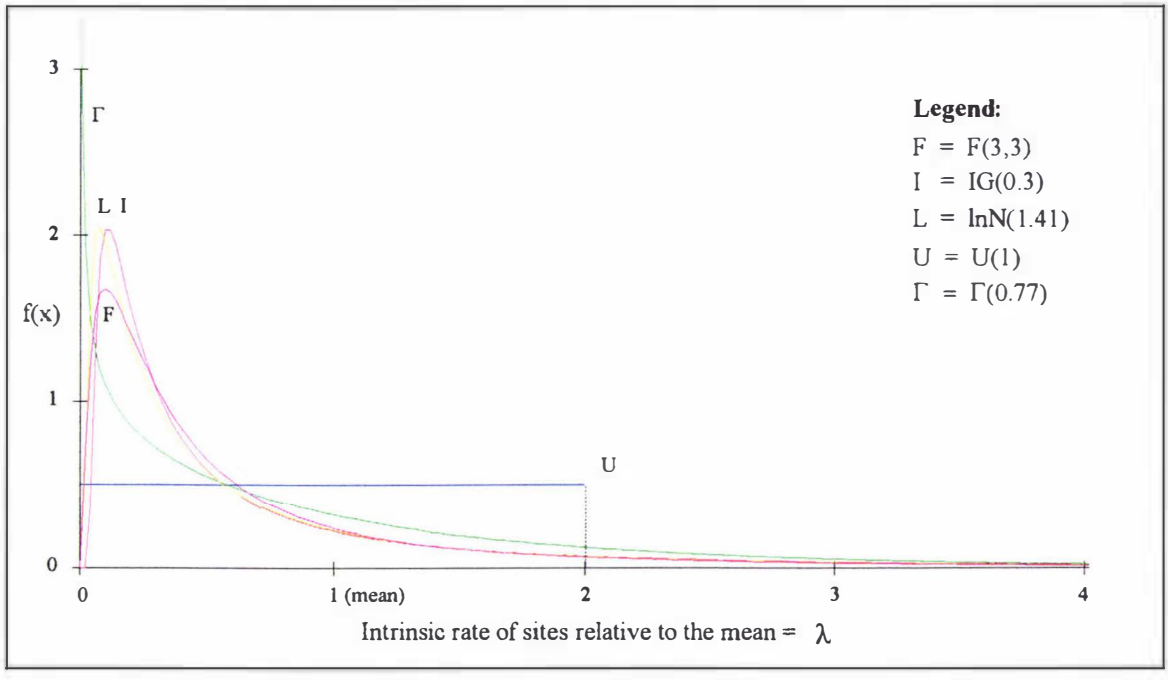


FIGURE 2.4 Shapes of different distributions of rates across sites. Parameters chosen are those that maximised the maximum likelihood fit of a 4-state model to sequence patterns reported in an alignment of conserved regions of rRNA (described further in text of fig 2.10, and chapter 5). All distributions have their mean set to one.

In contrast to the gamma distribution, the inverse Gaussian changes from the i.r. distribution to a progressively more skewed distribution (as shape parameter d decreases to zero) without the peak density shifting directly to the lowest rates of change (see figure 2.4 red line). Another feature not so evident from figure 2.4 is that when these two distributions (Γ and I) do have equal

amounts of their total density set below, say a relative rate of 5, the inverse Gaussian has the more slowly decreasing, so eventually greater, tail density (often only noticeable at values of λ = 10 or more times the mean rate). This in turn causes models based on the inverse Gaussian to tend to return higher values for the coefficient of variation of rates across sites than say the gamma distribution. In the next paragraph it is shown that this effect can make the c.v. a rather volatile statistic for characterising real data, even when the shape and explanatory power of two different distributions were quite similar. Nearly all of the sites in a sequence have relative rates below 4 times the mean, but the long tail to the right can have a profound effect upon the estimated c.v. despite there being few data points (sites) in this region.

Some distributions such as the lognormal or the F distribution do not have moment generating functions defined for all positive values, as some of their higher order moments (e.g. σ^2) do not always exist (are not finite). This is a consequence of their tails not falling away rapidly enough (sometimes called having flat tails). This same effect means that while, for example, over 99% of their density may lie below say λ = 10, and for $\lambda < 10$ their overall shapes are very similar (e.g. see fig. 2.4), their c.v.'s may be substantially different. For example, maximum likelihood fitting a distribution to some anciently diverged conserved region rRNA sites, suggested that the optimal k value for a gamma distribution was (c.v. = $k^{-0.5}$ = 1.14)(this data is cited in the caption to figure 2.10, and used again in chapter 5). However an inverse Gaussian distribution of rates across sites fitted nearly as well (only 0.5 lnL units difference) and gave d = 0.3, with a large c.v. i.e. $d^{-0.5}$ = 1.83, illustrating the point that the c.v. by itself is not such a useful summary statistic. As figure 2.4 shows, a lognormal fits closely to an inverse Gaussian for the majority of sites, however its inferred c.v. of 2.51 is nearly 40% bigger again when fitting to the same data. An F distribution also fits closely to the IG and Γ, but due to its tail being even flatter, it gives an even higher c.v. (for some values of the shape parameter the c.v. even becomes infinite). All the curves in figure 2.4 have more than 95% of their density below λ = 4 and more than 99.3% below λ = 10. The same effect can occur with "L" shaped distributions; for example the Weibull with shape parameter $c < 1$, can closely match the shape of the gamma distribution with $k < 1$, but has a more slowly decreasing tail density.

It is tempting to label sites with 10 or more times the mean rate as "hot spots", or hypervariable sites. Such terms are relative and not always appropriate. In some of the data we consider below, divergence dates are ancient (approximately 3 - 4 $\times 10^9$ (billion) years ago). Neutral substitutions in many organisms are fixed at a rate of approximately 10^{-8} per year, that is, we expect a mutation at any given site in about 10^8 years (100 million). Yet this is probably 40 or more times the average rate of substitutions in conserved regions of the 16S-like rRNA studied below. Thus most sites in the tails of the distribution of rates across sites for the conserved regions of 16S-like rRNA molecules are probably evolving considerably more slowly than neutral, making the term hypervariable somewhat misleading.

We will discuss further implications of this "tail" effect to models of molecular evolution elsewhere, emphasised here are:

(1) Within the region that contains most of the data (sites), the inverse Gaussian offers a close fit to both the lognormal and the F distribution's shape and should for most data give very similar results (with the exceptions noted below).

(2) When going from **r** to ρ then a "flat tailed" distribution may well imply substantially larger distances than a short tailed distribution of equal fit, because the sites characterised by the tails of these distributions are expected to have changed many times.

(3) Due to a distribution tail, the c.v. can be a rather non-portable statistic (i.e. varies between models) for measuring the implied variation of rates across sites, so we prefer to refer to distributions by their form and shape parameter(s).

Use of a lognormal, an F, or any other distribution which does not have a closed form moment generating function, requires a numerical integration to go from the true number of substitutions per site on a path or pathset to the observed number. To make a distance correction from observed to expected, then we must use the previous numerical integration method to construct a "lookup" table (or some similar method) for observed to inferred distances (e.g. Olsen 1987). When doing a numerical integration the tail of the probability density distribution needs to be truncated, and if this is chosen appropriately (e.g. $\lambda = 40$ would seem biologically cautious most of the time), then we would expect the inferences made with such a distribution (e.g. c.v. or number of substitutions along a path set) to be even more comparable to those inferred by the Inverse Gaussian. Distributions whose moment generating functions are defined, but not analytically invertible (e.g. the uniform), also require numerical methods to invert, but without need of tail truncation. The proof in appendix 2.2 shows that for the models we are considering here all $E[e^{X_l}]$ are finite and invertible (as long as all edges on the tree are of finite length), so all the integrals we need consider (closed form or numerically evaluated) do converge. An numerical integration could also be quite crude, for example, breaking a continuous distribution into ten equal sized classes, yet still approximately describe the distribution of rates across sites. One catch is that the more crude the integral, the less it tells us about the comparative fit of different distributions.

One distribution in table 2.1, the uniform distribution, does not have a tail and consequently has a bounded c.v., which with the mean set to one, has a maximum value of $\sqrt{(1/6)}$, or about 0.4. It seems very unlikely that the true distribution of rates across sites will conform to this shape. This makes it useful to compare with other distributions, and if the uniform fits about equally well as say a gamma, then we really don't have any clear idea of the shape of the distribution of rates across sites. The uniform distribution then helps to make testable the null hypothesis, "there is variation of rates across sites but just about anything with sufficient variability will fit about equally well" (in situations where the c.v. of the uniform is wide enough). Such a conclusion is expected to be only temporary until more data is obtained to refute the hypothesis, or a distinctly better model is applied to the data. A further step in the process would be to compare, for example, "triangle" shaped distributions (derived in appendix 2.4) with uniform distributions, and this would allow us to infer.more about the shape of the underlying distribution. A more direct approach is discussed later in this chapter.

2.3.4 Multi-modal distributions of rates across sites

It is also important to consider how to model more complicated distributions of rates across sites, including situations where there are multiple peaks (or modes). An important way to get multimodal distributions is as a sum or mixture of unimodal distributions with different means. This necessitates translating distributions, if sites are not going to be separated into classes. Translation is considered first, before moving onto the less general, but more tractable, case of a mixture of variable sites with invariant sites.

Translation is achieved with a linear transformation of the original variable, such that y = ax + b. In our case x has mean one, and b would be an amount added to reset the mean of y to one. When the mean of x is fixed to one, multiplying by a shifts the mean to $1/a$, which means that to reset the mean of y to one we add $(1-1/a) = b$. A well known statistical result (e.g. Stuart and Ord 1987, chapter 3) is the that after a linear transformation, y = ax + b, the moment generating function of y, is given as $M_y(t) = e^{bt}M_x(at)$ (an example of this type of translation is shown in appendix 2.4). Distributions of this kind might be envisaged for a collection of sites where we believe that there is a lower bound to the rate we expect sites to evolve at, and it is greater than zero. One such possibility is illustrated in figure 2.5 where we consider how the distribution of rates across sites might look if we did or did not translate the distribution at individual rate classes. (Compare figure 2.5a with 2.5b; note that this example is a mixed rate class distribution which means the degree of translation could vary for each component distribution).

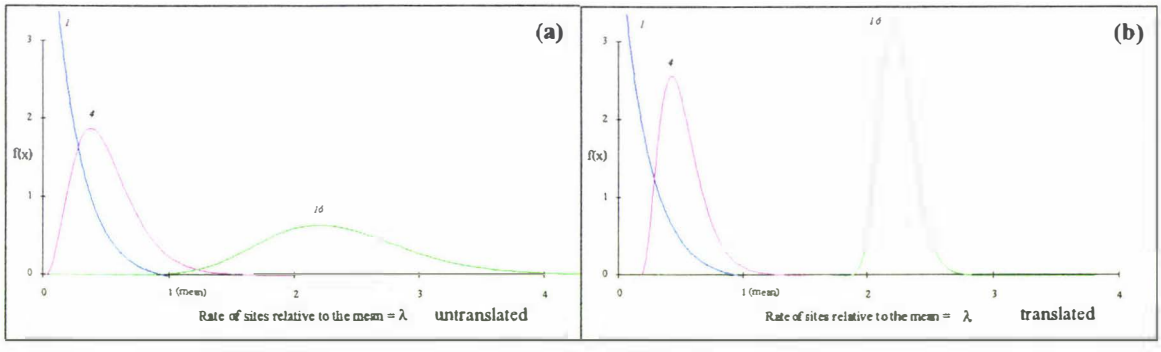


FIGURE **2.5a** Hypothetical distribution of rates across sites without translation of gamma distributions for different rate classes (mean over all classes fixed to 1). For example, a protein, where the second position sites have mean rate 0.25 (and shape 1), first position sites mean 0.5 (shape 4), and third position sites mean 2.25 (shape 16). **2.5b** A similar distribution with the same means and shape parameters, but with translation factors z = 0.25 for second and z = 1 for third position sites (see appendix 2.4).

Fortunately, when the sites are separated into their respective classes, altering the shape parameter of a distribution can closely approximate the overall effect of translation (for example in figure 2.5a the shape of the distribution for just third position sites in green, can appear like that of the translated third position sites in 2.5b by increasing the shape parameter, k, to about 100). This approach might give substantial error when the distribution was originally L shaped and then translated. Such a translated form of distribution is likely to be rare biologically since we would expect stabilising selection coefficients to steadily decrease and not cut off abruptly. If the situation should arise (e.g. a Γ distribution with $k = 0.7$ and the mean of 1 translated by 0.2 to 1.2) then a distribution like the inverse Gaussian might well offer a better fit than the

untranslated Γ. If we were doing maximum likelihood fitting of tree to data, then translation effects add extra parameters (and computation) to a model which already has unknown convergence properties (i.e. finding highest likelihood optima on a tree, Steel 1994b). However the computations can be done in reasonable time and in practice seem to converge well for many data sets. Hopefully the need for translation of standard distributions to achieve a significantly better fit of model to data will be rare.

We now look at ways of modifying Hadamard conjugations (and also most distance transformations, as we will explain in more detail in chapter 3) when distributions of rates across sites have multiple peaks or modes. Such distributions are expected to occur in protein coding regions, particularly due to the degeneracy of the genetic code. Fortunately there are some useful pathlength correction formulae which have closed forms even in this instance. In contrast to a single transformation for all sites grouped together, the alternative approach of grouping sites into rate classes requires a decision on how many classes there are to be, which sites belong to which class, what transformation will be applied to each class, and how the results will be recombined. The implications of grouping sites into rates classes are looked at in section 2.6.

With real sequences, the distribution of rates across sites is likely to be a mixture of distributions. This expectation follows from early work on the evolution of sequences. Fitch and Markowitz (1970) for example, argue that there are expected to be invariant sites in any coding region (as they code for amino acids at which any change would be lethal to the organism). Others such as Uzzel and Corbin (1971) have suggested that real data has a distribution of rates across sites well described by a gamma distribution.

Here we propose that it is reasonable to expect that both features predicted by Fitch and Markowitz (1970), and Uzzel and Corbin (1971) will hold in biological sequences. There will be both a proportion of sites (called p_{inv}) which are essentially incapable of accepting change in the period of evolution being considered, while those sites which can evolve will experience a range of either mutation rates (e.g. due to neighbor effects), or selective constraints (e.g. stabilising selection). These "variable" (i.e. able to vary) sites may have a distribution that is well approximated by a gamma or an inverse Gaussian distribution. We can analytically derive the moment generating function of this mixed distribution. For the invariant sites plus gamma distribution this gives,

$$r_i = p_{inv} + (1 - p_{inv}) \times ((k - \rho_i) / k)^{-k}, \qquad (2.3.4\text{-}1)$$

which has the inverse

$$\rho_i = k\,(1 - [(r_i - p_{inv}) / (1 - p_{inv})]^{-1/k}), \qquad (2.3.4\text{-}2)$$

while for the invariant sites plus inverse Gaussian we obtain,

$$r_i = p_{inv} + (1 - p_{inv}) \times \exp[d\{1-(1-(2\rho_i / d))^{0.5}\}], \qquad (2.3.4\text{-}3)$$

which has the inverse

$$\rho_i = 0.5d[1-\{1-(\ln[(r_i - p_{inv})/(1- p_{inv})]/d)\}^2], \qquad (2.3.4\text{-}4)$$

where p_{inv} is the proportion of invariant sites and k is the shape parameter of the gamma distribution, and d is the shape parameter of the inverse Gaussian distribution (see table 2.2).

More generally, we note that a proportion of invariant sites (p_{inv}) mixed with any other distribution of rates across sites λ, generates the new mixed distribution, λ', with moment generating function,

$$M_{\lambda'}(\rho_i) = p_{inv} + (1- p_{inv})M_{\lambda}(\rho_i), \qquad (2.3.4\text{-}5)$$

with inverse,

$$M_{\lambda'}^{-1}(r_i) = M_{\lambda}^{-1}[(r_i - p_{inv}) /(1 - p_{inv})]. \qquad (2.3.4\text{-}6)$$

Note that we have defined the invariant sites contribution in such a way that the mean of the sites able to vary is still fixed at one. Alternatively, to define the corrected pathset lengths to be the mean rate across all sites, then after making a pathset correction (observed to expected) simply multiply the result of equation 2.14 by the factor $(1-p_{inv})$. Such a model will often be less biologically appropriate since we usually want to disregard the invariant sites, and appreciate how much change has occurred amongst sites which are able to vary.

Except when treating a proportion of sites as invariant, then every time we add another component to our distribution of rates across sites we are adding at least two more parameters, the mean rate of these sites and their number, as a proportion of all sites. This results in more parameters to specify when going from ρ to $\mathbf{r}$. It also makes the analytic inversion of such equations very unlikely, although numeric inversion should pose few problems. In order to use equations such as 2.3.4-4 effectively, we require estimates of the associated parameters such as p_{inv}, k etc. In chapters 3 and 5 we describe a variety of ways we can make these estimates from the observed data. Lastly, we expect the invariant sites model plus a continuous distribution of rates across sites mixed model to offer a high degree of flexibility and robustness. As can be seen in figure 2.6 (in the next section 2.3.5), the p_{inv} parameter serves to shift the asymptote of the correction curve, while k alters its shape. Thus altering both together should reasonably well approximate the correction curve for any fixed distribution of rates across sites.

Apart from the case of invariant sites there are relatively few moment generating functions which have closed form inverses, unless we classify each site according to the rate of substitution. We can however use the moment generating functions of discrete distributions like the binomial or Poisson to specify invertible relationships. Drawing a line between the peaks of a Poisson distribution, produces a shape very similar to the Γ distribution, and such a distribution might find utility in testing for an periodicity in the distribution of rates across sites. In addition, collaboration with Dr Mike Steel revealed an analytically invertible trimodal moment generating function given in appendix 2.5. It works when the discrete rates of change fall into a pattern that allows the inverse of equation 2.3.3-2 to be solved by finding the roots of a quadratic equation, where the relative rates are specified as the coefficients. The equation may be useful when applied to sequences of coding regions, where for whatever reason the analyst does not wish to segregate them into coding position. Elsewhere we show how maximum likelihood via

Hadamard conjugations, can be used to quite directly infer the shape of multimodal distributions without placing restrictions on the overall shape.

The discussion here has focused upon pathset correction formulae for Hadamard conjugations. These same points are equally applicable to distance corrections for the same models. In chapter 3 we apply these same considerations directly to pairwise distance correction formulae.

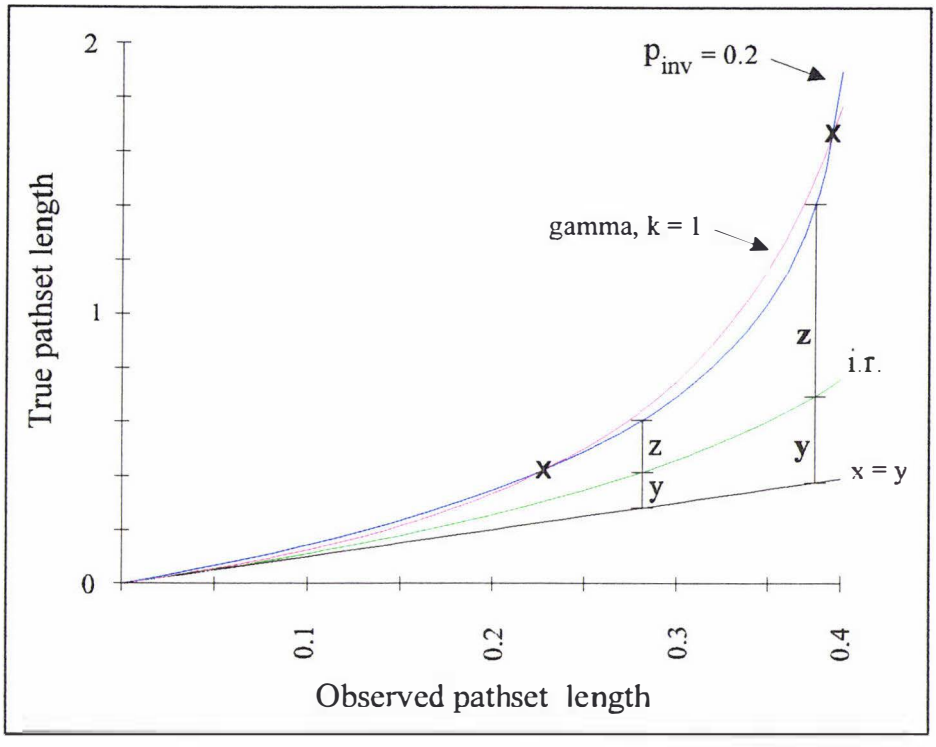


FIGURE 2.6 Path set correction curves with different distributions of rates between sites. The distributions are (1) identical rates (i.r.), (2) a gamma distribution with shape parameter k = 1, and (3) a bimodal distribution where a proportion (p_{inv} = 0.2) of sites are invariant with the remaining sites evolving at an identical rate. Note how two curves cross twice at true path lengths of approximately 0.4 and 1.8. They do not cross again as the invariant sites curve goes to infinity as the observed pathlength approaches 0.4.

2.3.5 Analysing transversional changes with extended Hadamard conjugations

This section gives examples of using extended Hadamard conjugations to "correct" 2-state data, such as purines (A or G) verses pyrimidines (C or T), for multiple hits. Figure 2.6 shows the relationship between observed and true path set lengths for; an identical rate of change at all sites; a gamma distribution of rates across sites; and a proportion of invariant sites. It is apparent that parameter p_{inv} (here set to 0.2) is shifting the vertical asymptote, whereas the gamma distribution with parameter k (set to 1 so giving an exponential distribution of rates across sites) is increasing the rate at which the curve rises over its whole length. Notice how in figure 2.6 the correction curves taking account of unequal rates at different sites are making approximately twice the amount of correction for multiple changes (the quantity marked y + z) that the standard i.r. correction makes (quantity y).

Just as "parsimony" applied to observed i.i.d. data is inconsistent (e.g. Hendy and Penny 1989), then a similar degree of inconsistency may occur when parsimony (or any other tree

selection criteria) is applied to a vector γ generated using the i.r. Hadamard conjugation if the sequences really evolved with a distribution of rates across sites (something illustrated later in this chapter). Importantly, even in such situations, the Hadamard conjugation still helps reduce the tendency for long edges to attract; it will still be making correction in the right direction, but it won't go as far as necessary. Lastly, notice how the invariant sites and the gamma distribution model can approximate each other over a substantial region. This significant ability of these two distributions (and many others such as the inverse Gaussian) to mimic each other suggests that we can gain a high degree of robustness even if we base our corrections upon the wrong distribution, as long as we estimate its parameters so as to fit the data as well as possible. This is a major theme of this thesis we will return to throughout as it has great utility in deriving robust methods for tree selection.

Figure 2.7 gives a biological example of the application of a Hadamard conjugation when taking into account a distribution of rates across sites. The data is a set of four 16S-like rRNA sequences from the data on archaebacterial phylogeny presented in Lake (1988), (using his alignment with all sites showing deletions in any taxa removed). We use the convention of hats (e.g. $\hat{\gamma}$) to denote estimators which are obtained from samples. It can be seen that varying the distribution of rates across sites according to either a gamma distribution or with a proportion of invariant sites makes a large difference. Assuming equal rates of change at all sites (the left of each graph), all three binary trees receive substantial support (figure 2.7 c and d). As we assume a greater spread of rates across sites support for two of these trees falls away, leaving only one candidate tree having support. This tree is compatible with what Lake (1986) calls the eocyte tree (excluding the position of methanogens).

The two distributions of rates across sites used in figure 2.7 give very similar results. If we were to plot the horizontal axis for the gamma distributed model as $1 / \sqrt{k} = k^{-1/2}$ (the coefficient of variation), rather than $1 / k = k^{-1}$, then figures 2.7a + c would look nearly identical to figures 2.7b + d (as is suggested by the similarity of the correction curves for each model as shown in figure 2.6). This is one of the pieces of evidence presented in this thesis to argue that to a first order of approximation, removing an appropriate proportion of constant (unvaried, but not necessarily strictly invariant) sites will largely compensate for the undercorrection that is generated by any distribution of rates across sites.

With the large amounts of change present in the data for figure 2.7, we can see evidence that, as we include a larger proportion of invariant sites (fig. 2.7 b and d), pathset values in the data are getting close to the vertical asymptote of the correction curves (see fig. 2.6) which leads to rapid change in γ values. For this data, when p_{inv} (the specified proportion of invariant sites) becomes $\approx$ 0.37 the argument of the logarithm in the correction formula becomes negative and the standard transformation no longer applicable. In addition, the variance of corrected distances increases as we assume either a lower value of k, or a greater proportion of invariant sites. Later in chapters 5 and 6, we see that these increased variances reduce the resolving power of this data considerably;

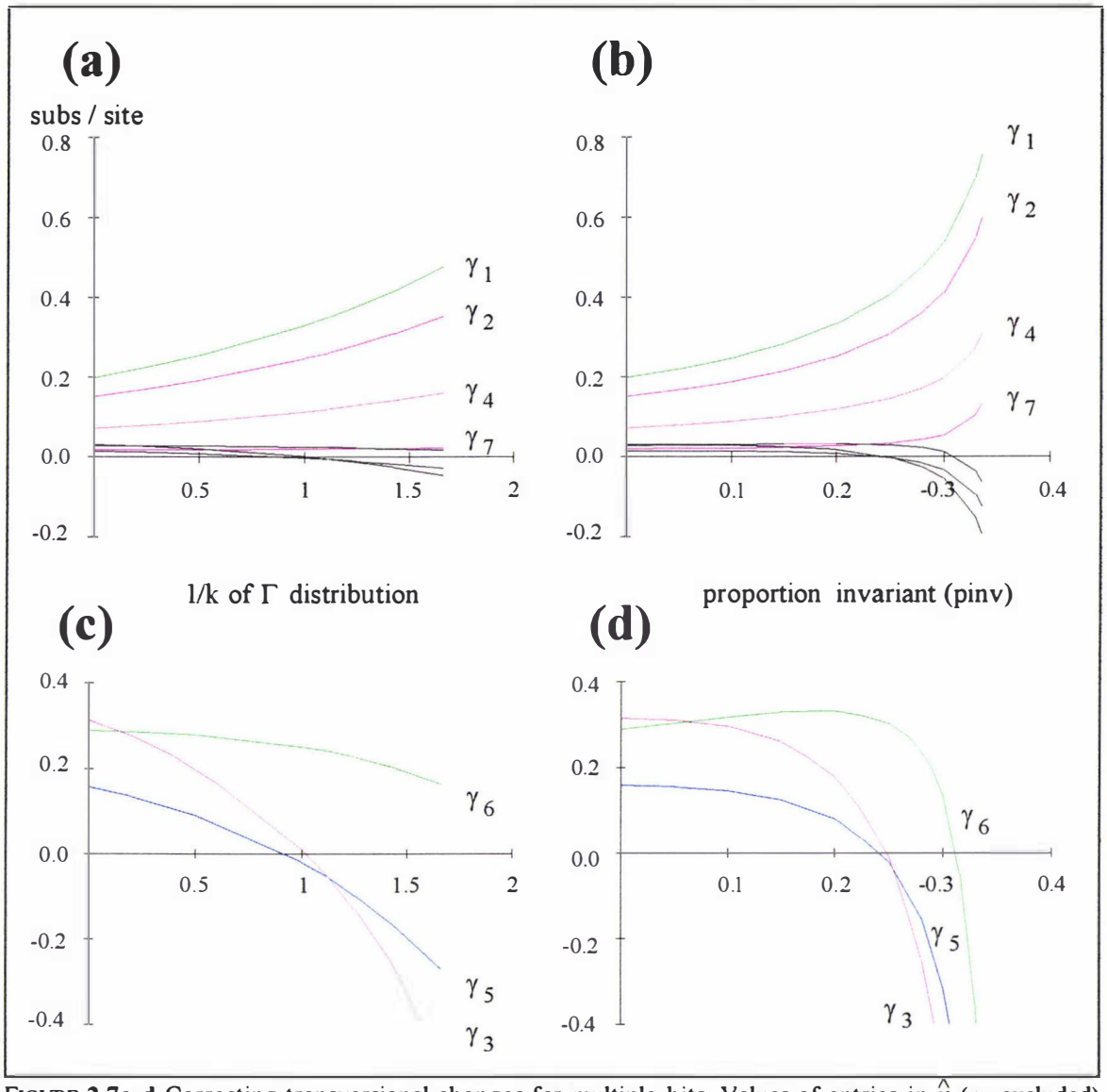


FIGURE **2.7a-d** Correcting transversional changes for multiple hits. Values of entries in $\hat{\gamma}$ (γ_0 excluded) after correction is made according to models that allow variation of rates across sites, applied to transversion patterns in the 16S-like rRNA of four taxa analysed by Lake (1988), with all sites showing any gaps excluded. The taxa are: (1) Eukaryote, *Homo sapiens*; (2) Eubacterium, *Escherichia coli*; (3) *Halobacterium salinarium*; (4) Sulphur metabolising bacterium, *Sulfolobus. solfataricus.* **2.7a + c** Corrected sequence patterns assuming a gamma distribution of rates across sites for various values of the shape parameter (its inverse is plotted on the x axis). **2.7a** All $\hat{\gamma}$ values, with just those pertaining to pendant edges labeled, while in **2.7c** just $\hat{\gamma}$ relating to internal edges on a tree are plotted and labeled . **2.7b + d**, as for **2.7a + c**, except the rates across sites are modeled as a proportion of invariant sites with parameter 'p_{inv}'. In both cases the apparent support for an archaebacterial tree (taxa 1 and 2 together, γ_3) begins high but then falls off steadily as the variation of rates across sites increases, leaving the eocyte tree (γ_6) as favourite. Both γ_3 and γ_5 have gone to zero at nearly the same point, a feature consistent with the data fitting the model. Nucleotide frequencies across all sites, and all taxa, in this edited data are A: 0.24, C: 0.25, G: 0.32, T: 0.19.This analysis provides a clear example, with real data, of the importance of being able to accommodate variation of rates across sites. Here we see evidence that the identical rates Hadamard conjugation is inconsistent when applied to this data. This inconsistency is to the extent that tree selection from i.r. corrected $\hat{\gamma}$ by methods such as compatibility, parsimony, or

closest tree results in a different tree to that which clearly best fits the data when rates across sites are taken into account. This is not a fault of the Hadamard conjugation per se, as the support for the wrong tree is even greater in the observed sequence data. Rather it illustrates the importance of the modifications made in this chapter.

an important fact ignored by many, including Gouy and Li (1989) who analysed similar data sets. A more extensive analysis of the data used by Gouy and Li (1989) is presented in chapter 3.

We suggest that as a general rule, many entries in γ will tend to be positive and not zero if there is insufficient correction for the multiple changes that have accumulated in the observed data (vector **s**). The situation in figure 2.3 illustrates this well. Here a prime condition for parallel changes, long edges (γ_1 and γ_2), juxtaposed with short ones, is apparently causing just this effect. It now becomes apparent that the adage of Hendy and Penny (1989) "long edges attract" is more general than originally described and can also apply to Hadamard conjugations when there has been insufficient correction for multiple changes between long edges. We will return to these features in chapters 5 and 6 where we also test the goodness of fit of $\hat{\gamma}$ to a tree, and discuss rules for diagnosing deviations from the model.

2.4 RATES ACROSS SITES 4-STATE HADAMARD CONJUGATIONS

We begin with a review of the essential features of 4-state Hadamard conjugations, before showing that they too can be extended to allow for variation of rates across sites. Following this we look at how insufficient correction for multiple changes on pathsets can cause 4-state Hadamard conjugations to be inconsistent and lead to the wrong tree being selected. We then analyse two sets of sequences and show the importance of pathset corrections which allow for unequal rates at different sites. The use of these extended Hadamard conjugations also reveals interesting patterns in the data.

In appendix 2.6 we describe an set of order 2^{t-1} Hadamard conjugations for 4-state data that are consistent under the same mechanisms of change as the order 4^{t-1} conjugations. In addition they can be extended to allow for a distribution of rates across sites, or restricted to the Kimura 2ST or Jukes-Cantor (Poisson) model. These were initially conceived independently of the full order 4^{t-1} conjugation (which was not fully developed until the second half of 1991) but may be considered special, but distinct, forms of it. They are kept separate from the rest of the thesis to help maintain the flow of other sections. How they fit in with other results in the thesis is described in appendix 2.6.4.

2.4.1 A review of the i.r. 4-state Hadamard conjugation

Székely *et al.* (1993) have described the conditions necessary for a Hadamard conjugation to be exact. As mentioned earlier the set of changes on the nucleotides must form a Boolean group structure. This structure may equally be described as the Abelian group "Z2 × Z2," which is also known as the "Klein four-group" with 4-states, which is a generalisation of Kimura's 3ST model

(Kimura 1981, Evans and Speed 1993). Given that all sites in a sequence on a specific edge in a tree evolve at the same rate according to the same Kimura 3ST rate matrix, then the model:

(1) Allows the rates of the three parameters in the rate matrix (transitions, and two types of transversions) to vary independently on each edge of the tree generating the data. i.e. each edge can have its own rate matrix as long as it conforms to the Kimura 3ST pattern.

(2) Is insensitive to the distribution of nucleotide frequencies at the root of the tree (i.e. does not require there to be an equal frequency of the four bases, but the model implies the sequences are evolving towards equal frequencies. We call models which predict equal base compositions, *equifrequency*).

This model may be called conditionally *nonhomogeneous* with respect to the transition matrix of character states on each edge of the tree, and conditionally *nonstationary* with respect to the base composition of each sequence (these terms are defined in section 1.4).

The vector **s** of nucleotide pattern frequencies has 4^{t-1} entries, and not 4^t entries because under Kimura's 3ST model four patterns of the 4^t possible sequence patterns always imply the same pattern of changes e.g. the patterns AAGG, GGAA, CCTT, TTCC are all equivalent indicating a transition between taxa 1 and 2 vs taxa 3 and 4. As with the two state Hadamard conjugation,

$$\mathbf{s}(T) = \mathbf{H}^{-1}(\exp(\mathbf{H}\gamma(T))),$$

where **H** is a $4^{t-1} \times 4^{t-1}$ symmetric Hadamard matrix which may be defined as the $2(t-1)^{th}$ Kronecker power of the matrix, $\begin{bmatrix} + & + \\ + & - \end{bmatrix}$, where + denotes 1, and - denotes -1.

Next is an explanation of the indexing which Székely *et al.* (1993) devised.

Table 2.3 Decoding entry $s_{i,j}$ into a nucleotide pattern (for explanation see following text)

i index	*j* index	*ij*	code	conv.	*ij* + conv.	new pattern
1	0	10	G	01	11	T
0	1	01	C	01	10	G
1	1	11	T	01	00	A
0	0	00	A	01	01	C

Following Székely *et al.* (1993), as also given in Hendy *et al.* (1994), then to deduce the nucleotide pattern of an entry in **s**:

(1) Index the whole vector from 0 up to $4^{t-1}-1$.

(2) Arrange the vector into a square matrix, by making a new row after every 2^{t-1} entries. Index the rows (*i*) and columns (*j*) of this matrix from 0 up to $2^{t-1}-1$.

(3) Give each entry in the matrix an index $s_{i,j}$, where *i* is the index of its row and *j* is the index of its column (if we were working straight from the vector form in 1, the index equivalently is $s_{i^*,j}$, where i* is $i \times (2^{t-1})$i). For example when working with 4 taxa (so the whole vector is

indexed from 0 to 63, and $2^{t-1} = 8$) the vector entry with index $s_{46} = s_{5*+6}$ (or row 5, column 6 in a table) which is written as $s_{5,6}$.

(4) Express each part of the index of s_{ij} as a binary number filling out to t places. Taking the i part first, our example gives 5 = 0101 (reading from the right), while for the j part, 6 = 0110.

(5) Transpose each part of the binary index to form a column, with the last taxon's state (the far left entry in each binary number) on the bottom of the column (see table 2.3).

(6) By convention we define here the index of each nucleotide state following alphabetical order i.e. A = 00, C = 01, G = 10, and T = 11. Use these indices to decode the lined up binary indices (ij) into nucleotide states for each taxa as shown in table 2.3.

(7) Since patterns are relative we wish to force any taxon to have a particular state (let's make taxon 2 to have state G instead of A) then deduce the required index to add modulo 2 (with no carry over in the last place) to achieve this transformation (here add 01). Add this conversion factor to the index of all taxa (again see table 2.3). Sometimes we will want to generate all 4^t sequence patterns, in which case we can generate each such permutation of a pattern by making the last sequence have an A (the standard indexing), C (add 01 to all patterns), G (add 10) and for T (add 11).

For all 3D illustrations of data with 4-state indexing within this thesis, the left axis gives the index j, while the axis to the right gives the index i of each pattern (see for example figure 2.8). When presented as a 3D plot, all entries not on the first row, first column, or the leading diagonal, show site patterns with 3 or 4 different states present.

When the frequency of nucleotides in the data are unequal, we can match the observed 4^t nucleotide pattern probabilities more closely with those predicted by the generalised Kimura 3ST model by assuming that the root of the tree had a nucleotide frequency that was not in equilibrium. If all tips of the tree are equidistant from the root then $\pi_{(l)} = \mathbf{P}\pi_{(r)}$, so $\pi_{(r)} = \mathbf{P}^{-1}\pi_{(l)}$, where $\mathbf{P}$ is the transition matrix from root (r) to tip (l) and π is a diagonal matrix of the nucleotide frequencies. This calculation is not accurate when it gives negative numbers for some entries in $\pi_{(r)}$, as this may occur if $\mathbf{P}$ is large and/or π is extreme. By the following calculations we can then deduce the expected frequencies of each of the four expansions of the 4^{t-1} patterns as $s_{i,j}*\pi_{(x)}$, where $\pi_{(x)}$ is the frequency of base (x) in the last taxon. While this procedure can help improve the fit of data to model predictions when judged over all 4^{t-1} patterns, it does not alter the fit when judged on the 4^{t-1} unique patterns of the equifrequency Kimura 3ST model (since $s_{i,j}$ = sum over x of $s_{i,j}*\pi_{(x)}$, since the $\pi_{(x)}$ must sum to 1).

Using this indexing, the first column of the square matrix expression of $\mathbf{s}$ refers to the proportion of transition bipartitions i.e. A and G or C and T states only in a column of data (in γ these entries refers to the rate of transitions on the corresponding tree edge), the first row to type 1 transversion bipartitions (A and C or G and T states only in a column) and the diagonal to type 2 transversions (A and T or C and G states). All other entries in the transformed data of γ (γ_0 also excluded) may be called model (non-tree) invariants, which should all converge to zero if the assumptions of the model are meet (Steel *et al.* 1993c).

2.4.2 Consistency of the extended 4-state Hadamard conjugation

As we have just seen the basic structure of the four state Hadamard conjugation is the same as that of the 2-state Hadamard conjugation. Since $\mathbf{H}\gamma$ is a linear orthogonal transformation and since $M_x(t)$ has the same properties as it does in the 2-state model (continuous and monotonic), it follows by the same arguments that the conjugations

$$\mathbf{s}(T) = \mathbf{H}^{-1}(M_x(t)(\mathbf{H}\gamma(T)))$$

and

$$\gamma(T) = \mathbf{H}^{-1}(M_x^{-1}(t)(\mathbf{H}\mathbf{s}(T)))$$

are also consistent (where $M_x(t)$ is the moment generating function of the distribution of rates of change across sites and $M_x^{-1}(t)$ is its inverse)(a mathematical proof is given in appendix 2.2). We give this model an explicitly mathematical development in Steel *et al.* (1993c), with respect to considering general conditions required for entries not in the tree to go to zero, i.e. be invariants. We will now consider how incorporating variations of rates across sites assists us in identifying the tree on which sequences evolved, how transition and transversion ratios may have changed through time in primate mtDNA, and what to make of the evolution of rRNA dating form the time of the common ancestor of all living organisms, approximately 3 to 4 billion years ago.

2.4.3 Unequal rates across sites causing inconsistency of tree selection

We now illustrate with some model-generated data, aspects of tree recovery when rates vary across sites. Hendy *et al.* (1992) used the four colour i.r. Hadamard conjugation to illustrate a case where parsimony applied to the uncorrected sequence patterns converges to the wrong tree (whereas application of the i.r. Hadamard conjugation recovers the tree in vector description). Here we show that taking the same model and allowing rates across sites to vary according to a gamma distribution, with $k = 0.4$, then the i.r. 4-state Hadamard conjugation (i.e. $k = \infty$) fails to correct the data enough to prevent tree selection criteria such as parsimony from being inconsistent. Further we show that commonly used tree selection methods applied to the i.r. γ vector, will chose the wrong tree. Figure 2.8a shows the observed pattern frequencies for our model. Next to them, figure 2.8b shows the data after application of the i.r. Hadamard conjugation. Three tree selection criteria (closest tree, compatibility and parsimony) will all choose a wrong tree, T_{13}, when applied to this "corrected" data. Notice however that in this example any of these criteria applied to just the transversions in either the corrected or the uncorrected data will identify the true tree (T_{12}).

Apart from selecting an incorrect tree, notice how much the identical rate corrections (figure 2.8b) have underestimated the rate of change on the long edges in the tree generating the data, making them out to be less than half their true length (0.4 + 0.1 + 0.1). This undercorrection camouflages the severity of the "long edges attract" problem. Consequently unless the transformation to account for unseen changes is accurate, it is difficult to diagnose accurately the severity of "long edges attract". It is interesting to note that in this example, closest tree (CT, Hendy 1989, Steel *et al.* 1992) or its close relative ordinary least squares (OLS, described in

section 5.2), both choose the wrong tree under more moderate conditions than unweighted parsimony or compatibility. By more moderate conditions we mean the ratio of long to short edges is decreased, or the variability of rates across sites is less, i.e. k larger. This is because both CT and OLS are minimising the sum of squares of edge lengths in choosing a tree (or conversely they are picking the largest sum of squared bipartition signals). This in turn leads to the largest erroneous signal (the multiple hit transitions) being given proportionately more weight with CT or OLS, relative to the linear weighting it receives when using compatibility. Accordingly, in this instance, the multiple hits in transitions lead CT and OLS astray before the methods of compatibility and parsimony.

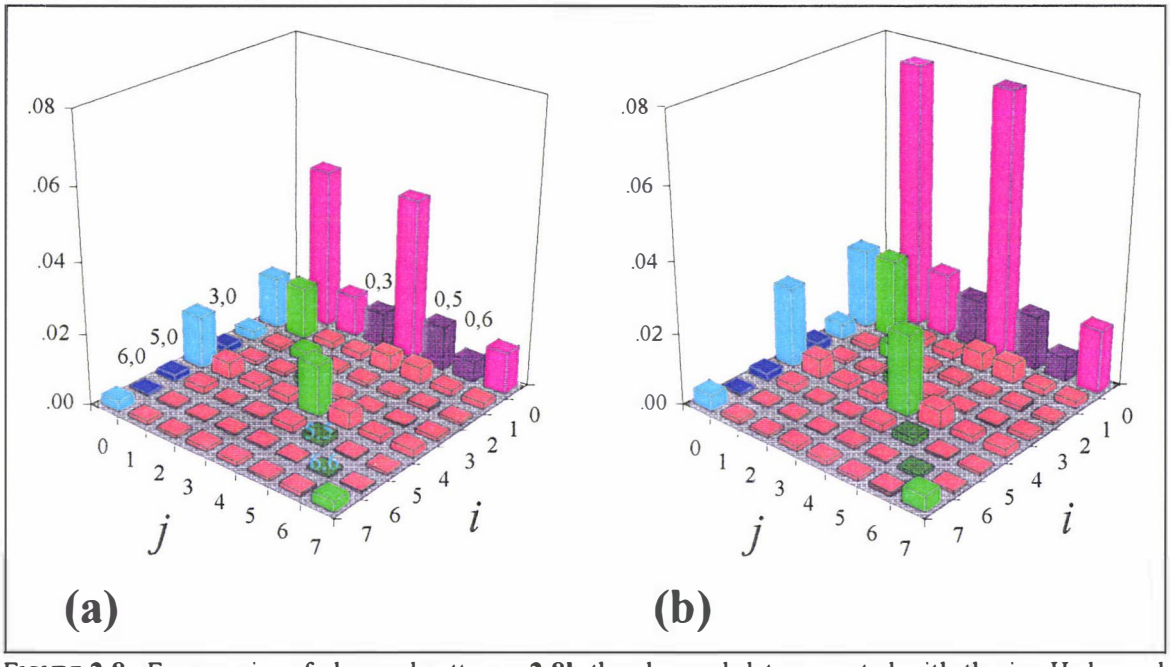


FIGURE **2.8a** Frequencies of observed patterns; **2.8b** the observed data corrected with the i.r. Hadamard conjugation (γ_0 excluded). **2.8a** The four state sequence spectrum for four taxa tree with a Γ distribution of rates across sites ($k = 0.4$). The tree generating this data is that used by Hendy *et al.* (1992)(tree T_{12}) to illustrate how parsimony applied to sequence patterns converges to a wrong tree (T_{13}). For the true tree, the rate of transitions (purple) on edges 1 and 4 are 0.2, on edges 2 and 7 are 0.04, and on the internal edge 3 is 0.02 (or indexing from 1 to 7, [0.2, 0.04, 0.02, 0.2, 0, 0, 0.04]). The rates of both types of transversions (tv1 blue, tv2 green) are the same and 1/4 the rates of transitions on the same edge. **2.8b** The gamma spectrum after application of the i.r. Hadamard conjugation to the observed patterns. Notice that the non-tree invariants (red) in the data are all of positive (if they were negative the graph would show a black square). The indexing is read right axis first (i), then left axis (j). Parsimony, compatibility or closest tree applied to these i.r. corrected sequence patterns will chose the incorrect "long edges attract" tree, T_{13}.

This example helps to demonstrate the need for "extra" correction for multiple changes when rates of change vary across sites, whether under the generalised transformation method of the Hadamard conjugation or with the more common distance matrix transformations. For this example we required a substantial (but realistic) coefficient of variation of rates across sites in order to get inconsistency of tree selection to show up in data corrected with the i.r. Hadamard conjugation. Here with 4-states, a more extreme model was required than would have been necessary with two character states and the same total edge lengths. The main reason for this is that the convergent and parallel transition patterns (the primary factor responsible for misleading

tree building methods in this example) are close to saturation, i.e. it took considerable heterogeneity of rates to increase their false signals enough that the partially effective i.r. correction did not negate their misleading effect. In addition sequence patterns showing both transitional and transversional change (red patterns) are not used by unweighted parsimony with four taxa, and hence cannot mislead it. These patterns can however help to mislead distance based tree building algorithms.

Having emphasised the general vulnerability of phylogenetic methods when neglecting the effect of unequal rates across sites, let us now highlight some advantages of using Hadamard conjugations even on data which violates the methods assumptions. Notice in figure 2.8b that the use of the i.r. Hadamard conjugation has improved the situation by decreasing the size of all non-tree signals despite the i.r. assumption not being meet. We expect this partial correction feature to be generally true for this method (although see section 3.4.3 which shows logarithmic corrections can sometimes make a situation worse). A second real advantage of using the Hadamard conjugation is that we can diagnose that something is going wrong. Just looking at figure 2.8b, there are obviously large positive incompatible signals amongst the bipartition patterns from which a tree is chosen using methods such as parsimony, closest tree and compatibility. Being able to see this is a clear warning that our method is probably in a region where tree selection could be inconsistent. Secondly, our suspicion that something may be amiss is reinforced by looking at the non-tree (model) invariants in figure 2.8b (i.e. those shown in red, all entries in figure 2.8b, except the top left and right rows and the diagonal heading straight down). That all of these entries are positive with many of them of apparently significant size, is a warning that the transformation is not correcting for a large proportion of multiple substitutions. This is the type of result that we also use to support the previous conjecture that with infinite sequence length non-tree bipartition patterns in $\mathbf{s}$ larger than expected under the model, will be greater than zero after the Hadamard conjugation. This example illustrates a real advantage of using Hadamard conjugations to analyse biological data. That is, Hadamard conjugations show the phylogeneticist the fit of data to model, sequence pattern by pattern and not just as a summary statistic.

While a few phylogenetic methods presently allow tests (e.g. evolutionary parsimony, Lake 1987) of the fit of model to data, the Hadamard method allows quick identification of the sequence patterns which may be causing a violation of the model due to the correspondence of $\mathbf{s}$ and γ patterns. This then allows diagnostic tests of the sequences to be run; for example are the sites relating to an anomalous pattern (say $\gamma_{4,1}$ in figure 2.8b) clustered spatially? If so then we might want to check for possible reasons, e.g in functionally correlated parts of the molecule under study, in an area of unusual base composition, or being near one end of a sequencing gel. This may in turn lead to resequencing or reediting as is deemed appropriate. Later in this thesis (chapter 6) we develop variety of statistical tests of the fit of model to data particularly with Hadamard conjugations.

2.5 SEQUENCE DATA ANALYSED WITH EXTENDED 4-STATE CONJUGATIONS

We will now look at the utility of using the rates across sites 4-state Hadamard conjugations to examine phylogenetic structure in nucleotide sequences. For this purpose we have chosen two data sets which are known to cause difficulties to commonly used methods, and potentially lead to the wrong tree being selected. In this section we use the term $\hat{\gamma}$, which as defined earlier in chapter one, is just the γ vector when it is estimated from sampled data (e.g. a biological sequence).

2.5.1 Analysis of 5kb of mtDNA relating to human origins

The first data set considered is a long unedited stretch of mtDNA from apes and humans (Horai *et al* 1992). It is expected this region will show a high degree of variability of rates across sites as it includes non-coding regions, protein coding regions (first, second and third positions) and regions coding for transfer RNAs (tRNA). In addition, mammalian mtDNA typically shows a very high rate of transitional to transversional changes, which can make both tree recovery and inference of transition to transversion ratios difficult. The frequencies of the nucleotides measured over all sites are $f_a = 0.302$, $f_t = 0.0.258$, $f_c = 0.308$, $f_g = 0.132$, which, assuming base compositions are at equilibrium, is an indication that they have not evolved via the generalised Kimura 3ST model (see section 1.9.2 for more detail on these sequences). However our model can closely approximate major aspects of their evolution since the last common ancestor (14 to 18 million years ago, see Waddell and Penny 1995), which is all we can expect any model to do. We do not claim that this is an ideal edit of the data (all sites grouped) for the purpose of estimating a tree. It is however interesting to study what occurs when we lump so many distinct regions (process partitions in the usage of Bull *et al*. 1993) together and examine the spectrum.

Figure 2.9 shows the results of our analysis using both a gamma distribution and an invariant sites model. We used both visual methods based on the tree-like appearance of the spectra (i.e. all entries not in the optimal HC tree having values close to zero) and maximum likelihood fitting (see chapter 5) to estimate the optimal parameters for our correction formulae. Both methods of estimation agreed closely and yielded $k = 0.35$ for the gamma distribution (fig. 2.9a,c,d blue columns), and $p_{inv} = 0.6$ for the invariant sites model (fig. 2.9a,c,d yellow columns)(compare with no correction red, or i.r. correction green). Either form of correction has made the spectrum of the transition bipartitions more tree-like in appearance, acting in particular to reduce the size of the signal for chimp-gorilla or human-gorilla, yet increasing the size of the signal for human chimp. This result implies that the apparent support for either a human-chimp or human-gorilla tree from this data is due to multiple substitutions. These transformations also imply that many more changes have occurred than are observed, and this applies to both transition and transversion events. In this respect the largest effect is seen with the invariant sites model (observations on other data sets suggests this holds generally).

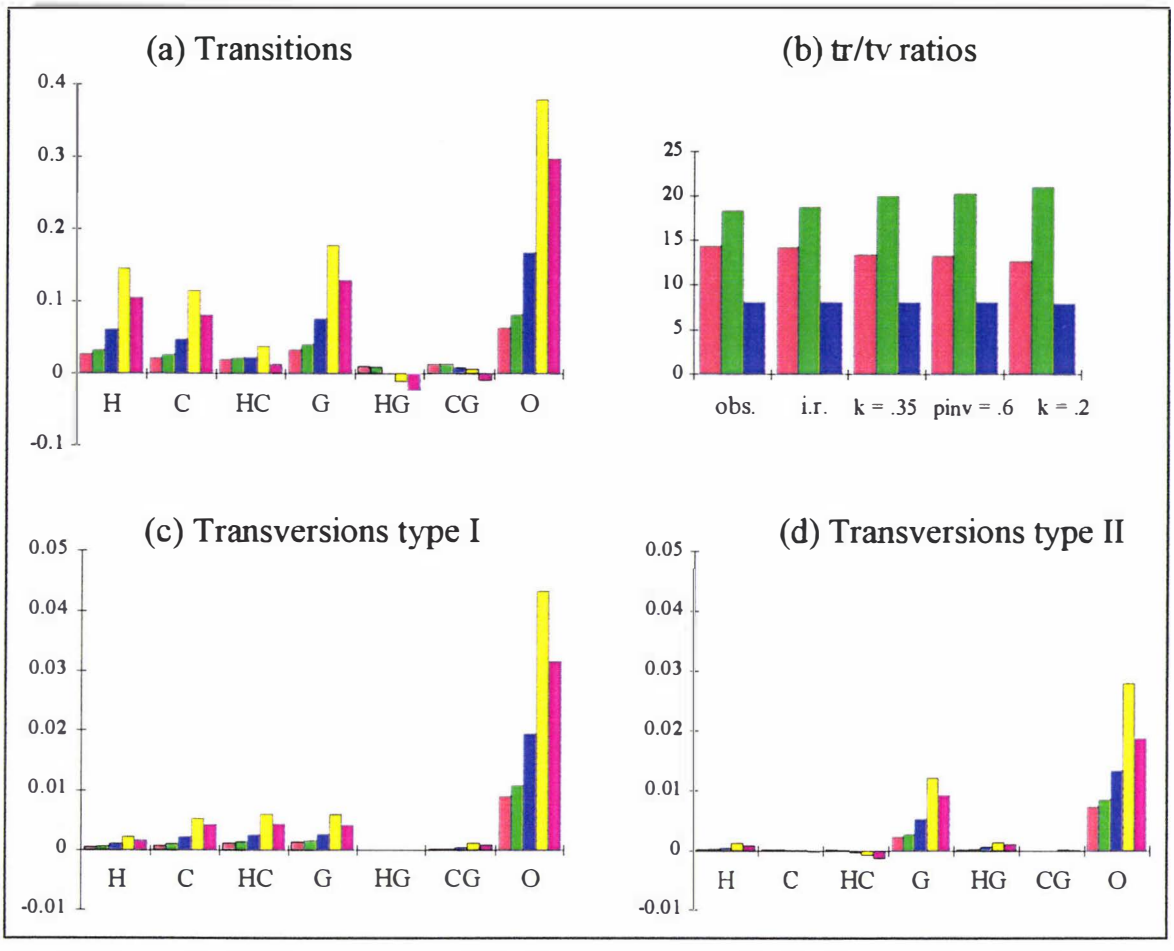


FIGURE **2.9a,c,d** The bipartition patterns of the four colour spectra for four hominoid mtDNA sequences (Horai *et al.* 1992) with corrections for different distributions of rates across sites, as follows; red observed signal size, green inferred by i.r. Hadamard conjugation, blue Γ with k = 0.35, yellow 60% of sites invariant, purple Γ k = 0.2. The ordering is "standard", while the labels are H = human, C = common chimpanzee, G = gorilla and O = orangutan. **2.9b** Transition to transversion ratios calculated by summing up the relevant bipartition patterns; first column of each set (red) is tr / tvI, the second (green) is tr / tv II, while the third (blue) is tr / (tv I + tv II). The correction made when calculating each set of ratios is shown under the x-axis, starting with the observed (uncorrected) pattern frequencies.

Included in figure 2.9 is an overcorrection for rates across sites (purple columns fig. 2.9a,c,d estimated with a gamma distribution, k = 0.2), as can be seen by certain values becoming distinctly negative (especially HG and CG signals figure 2.9a). Notice, however, that the apparent overcorrection has also resulted in a signal relating to the potential internal edge HG (in figure 2.9d) increasing in size, emphasising the non-linear nature of the corrections being made. Notice also how the size of the HC signal has decreased in size relative to the external edges in the transitional and type II transversional parts of the spectra, yet has maintained parity amongst the type I transversional changes. The patterns relating to sequence tri or quadra partitions (i.e. three or more states, not shown) are generally few per cell (between 0 and 5), although these are reasonably frequent when taken together (approximately 100). On this data the model invariants generally decrease in magnitude as k decreases (or p_{inv} is increased) until close to the point that the bipartition patterns in γ begin to show signs of overcorrection (i.e. become negative). The most frequent of these model invariant patterns have orangutan (the longest edge by far) being

separated from all other taxa by a transversional change, with the other taxa showing transitional changes amongst themselves.

A distinctive feature of this data is the high rate of transitions to transversions (the scale in fig. 2.9 c and d is approximately 1 / 10 that in fig. 2.9a). With observed data it is usual to calculate the ratio of transitions to transversions amongst a set of taxa as the sum of clear apparently single transition events (bipartitions) divided by observed transversion events (again bipartitions). This data gives an observed tr / tv ratio of approximately 8:1 (fig. 2.9b). This ratio remains nearly constant (fig. 2.9b) despite large differences in the corrections being made for multiple hits under some models. The observed rate of type I vs type II transversions is estimated to be approximately 4 : 3. With increasing amounts of pathset correction there is a slight trend towards inferring a higher proportion of the rarer type II transversions, and a slight fall off in transition / type I transversion rates (so the overall transition to transversion rate is nearly constant). The rate of transitions to transversions in the edge leading to orangutan appears to be lower than in the other taxa (note that this edge includes the edge from the root to the divergence of the three African apes). Whether this lowered tr / tv ratio is a real feature of the evolution of the orangutan, or an anomaly caused by the model underestimating the total number of transitions on this long edge remains unclear.

Transversions are relatively rare in these sequences, and consequently there are substantial statistical fluctuations expected in the proportions that occur (relative to the mean expected number). Most transversion bipartitions are represented by between 0 and 5 patterns in the observed data, although the long edge to gorilla shows about 10 transversions of each type, while the external edge to orangutan shows approximately 40. Later in chapter 6 we show how to test whether the difference in the inferred numbers of transversions amongst these taxa are significant, if so perhaps indicating a changing substitution parameters amongst even these closely related species. Lastly of the two classes of models analysed (gamma vs invariant sites) the invariant sites model could achieve a significantly better fit by likelihood than the gamma distribution (evaluations in Ch. 5), and visually this shows up as the relative size of signals supporting HC tree verses absolute size of those not supporting that tree.

Interestingly, the invariant sites model appeared to give the most tree like spectra with p_{inv} of about 0.6 (or 0.4 variable), which agrees well with the expected proportion of near neutral sites in this data (approximately 38%, made up mostly of third positions, some first positions, and short noncoding regions). Later in chapter 5 we use maximum likelihood to fit various models to an extended (more taxa) set of this data; finding that most models where predicting very close to 0.6 of all sites as invariant. The variances of entries in $\hat{\gamma}$ do increase as more sites are treated as invariant, and due to this factor using the appearance of $\hat{\gamma}$ as a guide to p_{inv} does tend to underestimate it. This inaccuracy is often removed when standard errors are recalculated taking account of the shortened effective sequence length, and the increased correction factors applied to all pathsets. A very rough and ready way of seeing this, is to consider the relative size of the signals not in the HC tree under the invariant sites model. It will later be shown in chapter 5 that

the optimal invariant sites model for this data does not have the resolving power apparent under the unrealistic i.r. model, none the less this data still strongly supports the HC grouping.

Lastly regarding some apparent lack of fit of data to model. One of the more outstanding non-tree signals is that for HG type II transversions (fig. 2.9d). The size of this signal is approximately 0.0015, and the adjusted sequence length (taking out invariant sites) is slightly less than 2000, implying that this signal amounts to 0.0015 × 2000 or just 3 events more than expected. It is not immediately apparent without knowing the variance and sampling distribution of this entry, whether this is a significant lack of fit. Later in chapter 6 we will discuss how to test if this pattern exceeds the size expected from stochastic error. There also appears to be a slight excess of transitions grouping CG. If a pattern should turn out to be significantly large, then it is possible to search along the sequence for where these patterns occur (e.g. do they predominate in degenerate first positions), highlighting the flexibility of using Hadamard conjugations to evaluate sequence evolution. In this case because excess transitions do not also occur supporting the HG pattern (which by symmetry of this nearly clock like tree should be about equal) suggests that they are probably a statistical fluctuation.

2.5.2 Analysis of anciently diverged rRNA sequences

Here we look again at the same sequences used to generate figures 2.7 (16S-like rRNA from human, halobacterium, "eocyte" and eubacterium) except using the modified 4-state Hadamard conjugation of section 2.4.2. The observed pattern frequencies are shown in figure 2.10a, and clearly the many substantial signals for site patterns showing three or more different states indicate that a high degree of divergence has occurred. The transformation chosen is Γ, with k = 0.77, since this distribution gave the highest likelihood (as evaluated by a tree building method described in chapter 5) amongst those compared. The inverse Gaussian (with shape parameter d = 0.30) gave a nearly identical fit (less than 0.5 log likelihood units worse), a proportion of invariant sites (p_{inv} = 0.28) gave a slightly worse fit (about 5 lnL units worse), but all were a substantially better (over 100 lnL units) than the i.r. model (see chapter 5 for more details of these analyses). Figure 2.10b shows the resulting spectra for the Γ model with k = 0.77 (the spectra for the other models looked very similar, but are not shown).

After application of the Hadamard conjugation the most immediately striking feature of the data is just how transitions and type II transversions have occurred on the external edges leading to human and *E. coli* (see figure 2.10b). We clearly expect a "long edges attracting" problem for estimating which tree has the most reliable support. After application of the transformation we see that many of the model invariants have been reduced, deviating about zero (both above and below, e.g. compare figure 2.10a with 2.10b). Some non-tree signals still appear comparatively large, particularly (7,1) and (7,4) (reading the right horizontal axis first). Furthermore, apparent rates (and ratios) of transitions and transversions vary substantially between lineages. Signals relating to internal edges on the corrected data are generally small, with the most prominent one being (3,3) a type II transversion signal supporting the eocyte tree, and a smaller transition signal supporting the halobacteria/eubacteria tree. Support for the grouping of archaeobacteria has

almost disappeared entirely, and this tree would, in addition, have to account for the deficit of transition and type I transversions if it were the true model. Even two external edge signals (2,0 and 2,2, transversions on edge to *Sulfolobus*) have gone to zero or less. Lastly, a most important feature that is often not noticed when doing phylogenetic analysis is how much larger the average signal is in the corrected data. The prominent tree signals in 2.10b are almost more than 2 times, with some up to 7 times, as large as in the uncorrected data (note the vertical axes differ in figure 2.10a verses 2.10b). Given such large amounts of correction via the Hadamard conjugation, we expect a nonlinear inflation of errors in the corrected data. This makes it difficult to judge how well this data really fits the model (and supports one tree over the others), without statistical estimates of expected stochastic error. Methods of estimating the expected magnitude of stochastic errors in $\hat{\gamma}$ are derived in chapter 4.

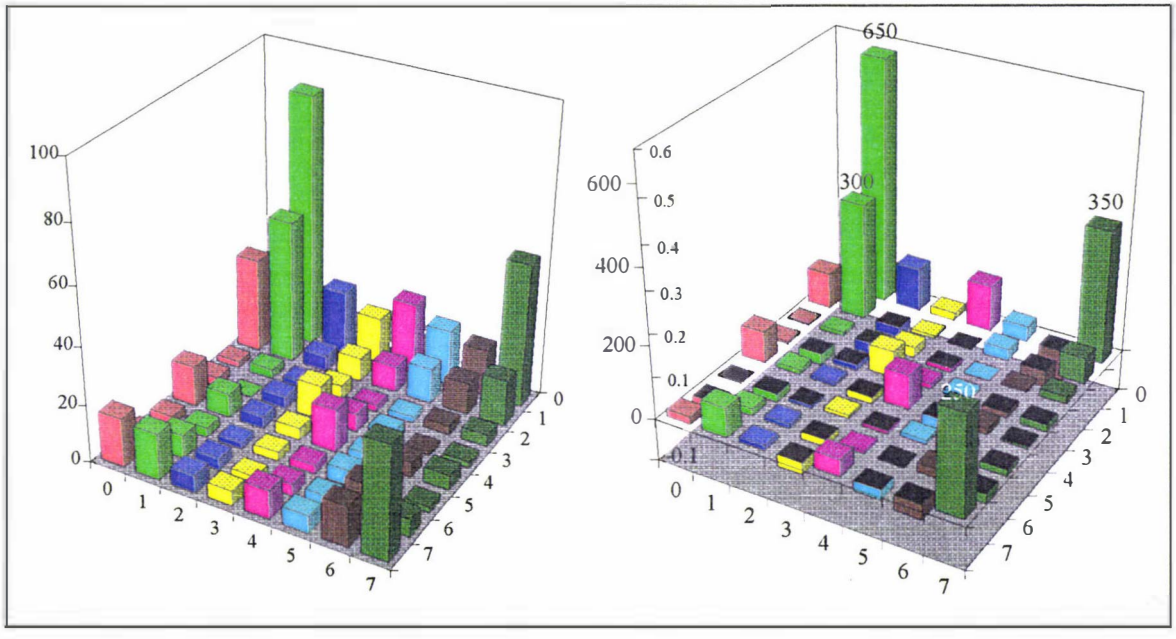


FIGURE **2.10** Correction of 4-state nucleotide patterns from conserved regions of 16S-like rRNA relevant to the earliest branchings of life (sequence length 1149). **2.10a** Observed patterns, **2.10b** patterns after the Hadamard conjugation with correction factor Γ ($k = 0.77$). The taxa and data are the same as those used earlier in figure 2.7, in order: eukaryote(1), eocyte(2), halobacterium(3), eubacteria(4). The vertical outside axis measures (a) the number of sites with each pattern, (b) the inferred weight of each pattern (numbers to left total inferred, to right divided by sequence length).

Without going into the intricacies of statistical testing, it is quite clear that just seeing this spectra should bring a note of caution when attempting to select a tree from this data. Ironically it is with just such data (especially deeply diverged rRNA) that we often see the most simple of phylogenetic methods applied. Typical examples are the work of Leipe *et al.* (1993) or Cavalier-Smith (1993) who use Jukes-Cantor i.r. corrections to very ancient 16S-like rRNA (sometimes applied to a mix of uncertainly aligned regions, leaving deletions in), and yet the authors appear to expect us to accept the results as substantial and statistically reliable, and sometimes even go on to define dozens of major new taxonomic groups based on such analyses! We suggest here that some of these authors would do well to peruse such data with Hadamard conjugations before giving their audience an interpretation of how reliable their results are. In actuality, this data

gives spectra of very similar appearance to other deeply diverged and aligned rRNA molecules. We should give pause to consider how reliable our methods are when applied to molecules so ancient that it is ridiculous to consider they are evolving by a neutral model, and indeed the very fundamentals of an i.i.d. model may be severely breached. Spectral analysis allows one way of evaluating the gap between our wishful thinking, and the "disrespect" which real molecules may show for our simplifying assumptions.

2.6 SEPARATING SITES INTO RATE CLASSES TO AVOID INCONSISTENCY

This approach can be viewed as an approximate way of apportioning sites to rate classes in order to give robustness to unequal rates across sites. More generally, grouping sites into distinct sets is a way to avoid inconsistency when groups of sites have evolved according to the same unweighted tree, but different processes (or "process partitions", in the sense of Bull *et al.* 1993). One variation of this general approach has been tried by Van De Peer et al. (1993) in the case of pairwise distances, with apparent success (although changes in the tree were was not evaluated statistically).

If it is known, or can be estimated which sites are expected to have different rates, then it is possible to separate these sites into distinct **s** vectors, transform each separately, then sum up the resulting gamma vectors. This provides an alternative strategy to attempting the maximization of the extra parameters in a multimodal distribution. This summed γ vector will be consistent if each of its constituent parts is consistent. We can call this an approximate method, due to its discrete nature when the real distribution may be continuous, and due to the uncertainty that is inherent in assigning a site to a rate class. However this term "approximate" tells us little about utility in practice, since, for example, we know that rates across sites do not follow simple mathematical distributions like the inverse Gaussian, or the gamma, yet we find these useful approximations. One benefit of separating sites into rate classes, is that it probably gives some robustness to inferring support for different trees when sites shift substitution rates relative to one another. Indeed this is possible when some sites encounter positive selection at certain times and evolve more rapidly during these periods. Another general example is provided by first, second and third coding positions in proteins where degeneracy in the genetic code allows for different amounts of neutral change at each site, but where positive selection will affect most first and second positions. (Note that even this basic allotment of sites into rate classes compensates for a lot of heterogeneity caused by stabilizing selection).

There are few established guidelines for separating sites prior to transformation, although it is common to separate sites into first, second and third coding positions, and various rRNA's (e.g. see Horai et al. 1992) even these classes often have very high degrees of rate heterogeneity. A better approach may be to be influenced by preliminary data analysis which can identify sites or regions of higher intrinsic rates of changes, unequal base composition, or possibly even regions which have evolved upon different (unweighted) trees, and have been juxtaposed by

recombination. In section 3.8, we consider separation of sites into different classes using the observed number of changes per site as a guide to which sites belong in which class. A similar approach can be applied to generate two or more **s** vectors of sites with similar intrinsic rates. Unfortunately, separating sites prior to transformation, then summing together, gives rise to a vector with larger stochastic errors then grouping first, followed by transforming all sites together (because of the non-linear nature of the pathset length transform). However, even this can be turned to an advantage, as we can estimate the variances of each set of transformed data and then weight them so as to minimise the overall variance of the summed γ vector. In doing this we are effectively placing most reliance upon those sites expected to be most informative in resolving the question at hand. The sizes of stochastic errors and some ideas on site weighting are considered in more detail in chapter 4.

2.7 DISCUSSION

In this chapter we have extended Hadamard conjugations to take into account variable rates between sites. It has been shown that such transformations are exact (consistent) under the model described, given the same assumptions of i.r. (identical rate) Hadamard conjugations, but allowing each site to evolve at a different intrinsic rate. A useful insight of this work has been that the integration of rates across sites to produce a consistent estimator of the true path length may be viewed as the inverse of the moment generating function of the distribution of the intrinsic rates of sites, allowing a useful link to be forged with known properties of statistical distributions. This has yielded closed form formulae for calculating sequence patterns in accordance with a uniform, triangular, gamma or inverse Gaussian distribution (this last distribution offering prospects of a good approximation to the lognormal, thereby reducing the need for numerical evaluations). We have also developed bimodal distribution correction formulae, that allow corrections of path lengths more in line with published predictions of how rates may vary across sites in functional molecules (see, Shoemaker and Fitch 1989). Our results showing how invariant sites may be simultaneously modeled with continuous distributions of rates across sites and answers Lake's (1988) desire for phylogenetic methods that allow this. We envisage that the methods developed here will be powerful in helping to answer the now 25 year old question of whether an invariant sites model, a unimodal or a multimodal distribution of rates across sites best describes the overall pattern of stabilising selection in a stretch of DNA (Fitch and Markowitz 1970).

The question of what sort of distribution best characterises real DNA sequence rate variation in specific data sets is largely untested. Our expectations at this point are that protein coding molecules should show some trimodal character due partly to the varying degrees of degeneracy of the genetic code for the first, second and third positions. Different selection pressures on amino acid sites could also result in each of these three classes having a characteristic distribution of rates across sites. We expect that these distributions will be most skewed (i.e. a high c.v.) at first and second positions. It should also be expected that in some regions there will

be many sites which are at least temporarily unable to vary, i.e. are invariant (Fitch and Markowitz 1970). This model is considered further in chapter 5. Better understanding of the distribution of rates across sites in a molecule should help avoid indeterminacy in which tree the data is best supporting.

While characterising the distribution of rates across sites is a practical problem, a more interesting and challenging problem is investigating the causes of unequal site rates. It is commonly accepted that stabilising selection acts upon sites to reduce rates of substitutions, but little is known of the actual selective coefficients, how constant they are through time, and whether they interact or correlate with particular processes occurring at particular sites, or whether these are constrained by more global selection pressures on the whole molecule. Indeed the more we consider i.i.d. models, the more clear it becomes that sites which show a substitution rate significantly slower than neutral substitutions (yet may still show synapomorphies amongst present day taxa) must be evolving by some sort of covarion model. That this must be so is very simply demonstrated. Consider a site which is evolving more slowly than neutral, the immediate thought is to say that it fits an i.i.d. model, but that it has stabilising selection acting upon it. However while such a site is less prone to undergo substitution, once it has substituted there is nothing to stop it from mutating back to its previous state which must still have a higher selective advantage. There is no way that this sort of process can explain the historically informative substitutions that are evident in molecules with divergences many times older than the "half-life" of neutral substitutions. Hence if a site is to evolve markedly more slowly than a neutral site and be phylogenetically informative of ancient divergences, then the state it changes to must become selectively optimal, just as we assumed the previous state was. Unless we digress to a "Red Queen" type of hypothesis (e.g. Nei 1987), the best explanation of this type of evolution is concomitantly variable sites (the covarion model for short, Fitch and Markowitz 1970).

It will be interesting to see if under a benevolent covarion model it is possible to show that the expected sequence patterns will still converge to some fixed set of probabilities (e.g. Bernstein's theorem, Rényi 1970). Consequently our i.i.d. models may still be reliable estimators for some covarion mechanisms of evolution. However other covarion models can potentially be very misleading, such as when there are unequal numbers of invariant sites in different lineages (as detected in Lockhart *et al.* in press).

While it is has generally been assumed that additive distances guarantee recovery of the true tree, this need not be the case. In chapter 5 we encounter examples where different trees give the same sequences. Steel *et al.* (1994) have shown by an existence theorem, that different rates of change across sites can result in distinct binary trees giving identical sequences. Chapter 5 shows that even more commonly, unresolved trees and binary trees give the same sequences, or two different binary trees give the same pairwise distances. While later it is also shown that the random addition of extra taxa should break these tie situations of different trees giving the same sequences, their occurrence does highlight one way models can be comprehensively fooled with limited data. Thus it is desirable to learn more about the true distribution of rates across sites, something which not only facilitates identification of the unweighted tree, but also will assist in

obtaining better estimates of edge weights. Accordingly in this thesis we identify two types of additive; generally additive (just called additive) which does guarantee consistency, or locally additive (i.e. additive only on certain weighted trees) which does not.

The theory and the correction formulae developed here apply directly, or with little modification, to many pairwise distance correction formulae and their underlying models. One example is a the i.r. time reversible correction formulae (e.g. Lanave *et al.* 1984, Tavaré 1986), which like the Hadamard conjugation, takes the logarithm of a vector in an intermediate stage of the calculations. In chapter 3 we extend this general distance to allow for a distribution of rates across sites, which can use any inverse moment generating function to make corrections.

Hadamard conjugations for calculating pattern probabilities, or likelihoods, can be coupled to a numerical optimisation routine to find maximum likelihood solutions on a given tree, and then search across trees. These calculations were used extensively to study DNA sequences from apes and humans in 1993 for the book chapter Waddell and Penny (1995), which was accepted and to have been printed in that year (see chapter 5 for details of these calculations). In the same year Yang (1993) used numerical integrations to also illustrate likelihood calculations when rates vary across sites.

Application of extended Hadamard conjugations to real data has shown promising results. In our first example, using just transversional changes, seeing $\hat{\gamma}$ cautioned us away from accepting results that would be obtained from tree selection applied to an i.r. model. Instead the results pointed towards a still controversial description of the origins of the oldest lineages known (eukaryotes, eubacteria and archaebacteria), the so called "eocyte" tree (Lake 1986) (the significance of this result is determined in chapters 5 and 6). Application of the extended 4-state Hadamard conjugation, however, showed a more complicated situation. These analyses highlighted the need for reliable statistics upon Hadamard conjugation results (and by implication any method of phylogenetic analysis). Analysing ancient rRNA lineages begs the fundamental question of whether even the most general methods currently employed are giving an adequate fit of model to data (we develop such tests in chapter 6). Note that the need to optimise parameters related to the fit of distributions of rates across sites is also forcing this issue, since the fit of data to model is our most direct criterion for deciding on the form of these distributions.

Appendix 2.1 Proof of the inconsistency of Hadamard conjugations if sites change their relative rates

We demonstrate, by way of a counter example, that Hadamard conjugations applied to data with a distribution of rates across sites need not be consistent if sites change their relative rates. Using the Hadamard conjugation to predict sequence pattern probabilities shows that in this situation maximum likelihood methods applied to sequence data will also make errors.

Let the classes of sites have two different rates, λ_1 and λ_2, with proportions $p_1 = p_2 = 1/2$. Sites in one class exchange rates with those in the other class after t_1 units of time, and remain like this for a further t_2 units of time. After (t_1+t_2) units of time, then by equation 2.3.3-2

$$r_i(s) = 1/2\exp(-2(t_1\lambda_1+t_2\lambda_2)) + 1/2\exp(-2(t_1\lambda_2+t_2\lambda_1)) \neq \qquad \text{(A2.1-1)}$$

$$1/2\exp(-2(t_1+t_2)\lambda_1) + 1/2\exp(-2(t_1+t_2)\lambda_2) = r_i(c)$$

where (s) denotes sites swapping their relative rates and (c) denotes sites keeping constant relative rates, and the factor -2 is a result of changes per site (in γ) being converted to this form (in ρ) by the Hadamard transform. While in both cases the true number of substitutions is equal on the designated path, $r_i(s)$ is always less than $r_i(c)$, suggesting that the **s** (sequence) vector for the shifting rates model will show more divergence (Note: $r_i(c) - r_i(s) = 1/2(A^{t1} - B^{t1})(A^{t2} - B^{t2}) \geq 0$, where $A = e^{-2\lambda 1}$, $B = e^{-2\lambda 2}$, with equality <=> $\lambda_1 = \lambda_2$). Since $\mathbf{s} = \mathbf{H}^{-1}\mathbf{r}$ is an orthogonal invertible transformation, this implies that **s**(c) must be different to **s**(s), with **s**(s) showing different pattern probabilities. Conversely with sequences generated under the (s) model, then ρ(s) will always be greater than ρ(c). For pathsets with 4 or more end points the same inequalities also hold. This inability to estimate pathlengths applies equally to distance estimates, making them inexact also.

The Hadamard conjugation can be used to explore the effects of such shifting site rates (analogous to a generalised covarion model) upon phylogeny estimation, by generating the expected sequences under Cavender's model. To do this estimate entries in **r** using equation A2.1-1, taking parameters for times in different rate categories on a tree. Perhaps a simpler way of seeing this is to take all the sites in set one which are evolving identically, according to $T_1(w_a)$, (tree one, with edge weight set a), giving sequence vector $\mathbf{s}_{1a}$. Those sites in set two are evolving according to tree $T_1(w_b)$ (same tree, different weight set), giving vector $\mathbf{s}_{1b}$. The overall sequence vector, $\mathbf{s}_{1ab}$ is just the weighted average of $\mathbf{s}_{1a}$ and $\mathbf{s}_{1b}$ (the weighting being the proportion of sites in each set). Note that $\mathbf{H}^{-1}\ln\mathbf{H}(\mathbf{s}_{1a} + \mathbf{s}_{1b}) \neq \mathbf{H}^{-1}\ln^{-1}\mathbf{H}(\mathbf{s}_{1a}) + \mathbf{H}^{-1}\ln\mathbf{H}(\mathbf{s}_{1b})$, since generally $\ln(a + b) \neq \ln(a) + \ln(b)$. Consequently under this mixed site rate model, the Hadamard conjugation will be inconsistent in not recovering $\gamma = \gamma_{1a} + \gamma_{1b}$. It is even possible that under some circumstances γ could deviate from expectations enough that tree selection procedures applied to γ would give an incorrect tree. It would be interesting to consider the reliability of current methods of tree selection under this type of model, perhaps most easily with simulations. In particular, it would be good to know if using a wide class of correction formulae (M^{-1}), with some criterion for picking that which maximises the treeness of γ, could substantially improve robustness with

these types of sequence evolution. This approach should be able to be extended to offer a useful way to derive the sequence pattern probabilities of other covarion type models, and hopefully a way to begin to understand them.

A recent paper by Penny *et al.* (1994) explores a model where each site evolves by the 2-state Poisson process, yet its intrinsic rate is unknown and varies from edge to edge independently of all other sites. This model (with no other constraints applied) has as its maximum likelihood solution unweighted parsimony applied to the observed sequence pattern probabilities. This model maybe considered an extreme case of the model described here. It will be interesting to see for which intermediate models, a consistent estimator of the tree (recovering edge weights or not) can exist.

Appendix 2.2 Proof of equation 2.3.2-4: The consistency of extended Hadamard conjugations

Let the rates (λ_j) of sites j in a sequence be randomly drawn from probability density function, $f_x(x)$, which has moment generating function $M_x(t)$, = M, let the process of evolution at a site be one for which the consistency of Hadamard conjugations has already been proven (Steel *et al.* 1992, Hendy and Penny 1993), and let the number of each type of change along an edge at site j, $\gamma_\theta(j)$, be written in the form $\gamma_\theta \times \lambda_j$ (here θ is just the index of entries in the frequency vectors **s** and γ, and c is the sequence length).

Then the Hadamard conjugation, $\mathbf{s} = \mathbf{H}^{-1}\mathbf{M}\mathbf{H}\gamma$ is consistent, and invertible to give $\gamma = \mathbf{H}^{-1}\mathbf{M}^{-1}\mathbf{H}\mathbf{s}$ (for two, four ,or any power of two, of states).

Proof:

Let $\gamma = [\gamma_\theta]$ where $\gamma_\theta = q_\theta$, if $\theta \in \varpi(T)$ (the set of weighted edges, q_θ, comprising the tree), while $\gamma_\theta = 0$ otherwise (excepting γ_0 which is $-\Sigma\, q_\theta$). Then the expected value of $s(j, \theta)$, denoted, $\langle s(j, \theta)\rangle$, is given by theorem 1 of Steel *et al.* (1992) as,

$\langle s(j, \theta)\rangle = (\mathbf{H}^{-1}\exp\mathbf{H}(\lambda_j \gamma))_\theta$. (this result is for the 4-state Hadamard conjugation, while Hendy and Penny (1993) give an equivalent result in describing the 2-state Hadamard conjugation).

Thus, $$\frac{1}{c}\sum_j \langle s(j,\theta)\rangle = \frac{1}{c}\sum_j (\mathbf{H}^{-1}\exp\mathbf{H}(\lambda_j\gamma))_\theta ,$$

$$= \left(\mathbf{H}^{-1}\frac{1}{c}\sum_j \exp\mathbf{H}(\lambda_j\gamma)\right)_\theta$$

Now,

$$\lim_{c\to\infty}\frac{1}{c}\sum_j(\exp\mathbf{H}(\lambda_j\gamma))_\theta = \lim_{c\to\infty}\frac{1}{c}\sum_j(\exp\lambda_j\mathbf{H}(\gamma))_\theta = \lim_{c\to\infty}\frac{1}{c}\sum_j e^{\lambda_j(\mathbf{H}\gamma)_\theta} = M((\mathbf{H}\gamma)_\theta)$$

Thus $\lim_{c\to\infty}\frac{1}{c}\sum_j < s(j,\theta) >= (\mathbf{H}^{-1}M\mathbf{H}(\gamma))_\theta$,

If we can rearrange pairs of sites (i,j) in descending order of correlation (of evolutionary change), such that the correlation between sites eventually falls away as rapidly as $\frac{1}{|i-j|}$, then applying Bernstein's theorem (see Rényi 1970, p. 379), we deduce that for each θ,

$$s_\theta \to_p \lim_{c\to\infty}\frac{1}{c}\sum_j < s(j,\theta) >= (\mathbf{H}^{-1}M\mathbf{H}(\gamma))_\theta ,$$

thus $\mathbf{s} \to_p \mathbf{H}^{-1}\mathbf{M}\mathbf{H}\gamma$. (where $\to_p$ denotes convergence in probability) (A2.1–1)

Now because M is continuous, then so is $\mathbf{H}^{-1}\mathbf{M}\mathbf{H}$. In general for random vectors $Z\to_p z$, so it is easily confirmed that $\phi(Z)\to_p \phi(z)$ for any continuous function φ. Application of this to A2.1-1 gives the desired result, such that $\mathbf{H}^{-1}\phi\mathbf{H}\,d \to_p \mathbf{H}^{-1}\phi\mathbf{H}\mathbf{H}^{-1}\mathbf{M}\mathbf{H}\gamma = \gamma$.

Appendix 2.3 Deriving moment generating functions while fixing the mean to one

Let the X be the random variable relating to the intrinsic rate of site *i* (more generally called λ). Assuming X to be drawn from a uniform distribution, with mean fixed to 1, and obeying the requirement that X is always positive, then the probability density function,

$f(x) = \begin{cases} 1/(2b), (1-b\le x\le 1+b) \\ 0, elsewhere \end{cases}$, $0 < b \le 1$, where parameter b measures the spread of the X's. Then the moment generating function is derived as

$$\begin{aligned} M_X(t) = E[e^{Xt}] &= \int_{1-b}^{1+b}\exp(tx)f(x)dx \\ &= \int_{1-b}^{1+b}\exp(tx)(1/2b)dx \\ &= \frac{1}{2b}\int_{1-b}^{1+b}\exp(tx)dx \\ &= \frac{1}{2b}\left[\frac{1}{t}\exp(tx)\right]_{1-b}^{1+b}dx \\ &= \frac{1}{2bt}\left[\exp(t(1+b))-\exp(t(1+b))\right] \\ &= \frac{1}{bt}\sinh(t(1+b)) \end{aligned}$$

For our purpose here we equate t with the path set length (ρ_i), while $M_X(t)$ we equate with the observed pathset length r_i, i.e. $r_i = (bt)^{-1}\sinh(\rho_i(1+b))$(note in this section "t" is not the number of taxa). We can also use the moment generating function for its more usual purpose of deriving the s.d. of our underlying uniform distribution. Differentiating twice and setting t to zero we obtain

$E[X^2]$(not shown), so s.d. = $\sqrt{E[X]^2 - E[X^2]}$ = $\sqrt{2b/12}$ or $\sqrt{b/6}$ (which is also the coefficient of variation since μ is set to 1). Unfortunately the inverse of this moment generating function does not have a closed form, since t appears in two noncombinable exponential expressions.

We may perform a similar integration for the other normalised distributions. One approach is before performing the integration, set the mean of the underlying p.d.f. to one, by replacing the random variable x with ($x\mu$) and then multiplying the whole p.d.f. function by μ (i.e. $f(y) = \mu f(x\mu)$)(where μ is the mean of the p.d.f. of x). The variance of the underlying distribution with mean set to one, f(y), then becomes $(1/\mu)^2 var(x)$.

If we have their moment generating function from the literature, then simply replace t by (t / μ), in order to set the mean of underlying rates across sites distribution to 1 (where μ is the expected value of the random variable X with mean not equal to 1). This holds since if Y = X/a, then $M_Y(t) = M_X(t/a)$. For example, the standard gamma distribution has mean (k), and variance (k), (Stuart and Ord 1987, p 192), $M_X(t) = (1-t)^{-k}$, so setting the mean of the underlying distribution to one we get $M_Y(t) = (1-t/k)^{-k} = ((k-t)/k)^{-k}$ as shown in table 2. Our rescaled underlying gamma distribution of rates across sites now has mean 1, and variance $1/k^2 \times k = 1/k$, so the c.v. = $1/\sqrt{k} = k^{-0.5}$.

Note that the characteristic function (c.f.) of a statistical variable = $E[e^{ixt}]$ is often given in place of $M_X(t)$. We may convert it to the moment generating function by replacing t with -it (where i is the imaginary number), which equates to simply dropping the i's from the c.f. {i.e. $c.f._X(t) = E[e^{iXt}]$, $c.f._X(-it) = E[e^{i(-it)X}] = E[e^{Xt}] = M_X(t)$}. For example $c.f._X(t) = (1-it)^{-k}$. (Stuart and Ord 1987, p. 192), so $M_X(t) = (1-t)^{-k}$.

Special thanks to Terry Moore for assistance when deriving this distributions moment generating function.

Appendix 2.4 The moment generating function of translated distributions

We begin this example by deriving the moment generating function of a standard triangular distribution (with p.d.f., $f_x(x) = 2x$, $0 \leq x \geq 3$), then translate this to turn it around so that it points in the direction of higher rates, while simultaneously fixing its mean when allowing its base length to change.

$f_x(x) = 2x$, $0 \leq x \leq 1$, 0 elsewhere.

$$M_x(t) = E[e^{tx}] = \int_0^1 2xe^{tx}dx$$

$$= 2\left[\frac{x}{t}e^{tx}\right]_0^1 - 2\int_0^1 \frac{1}{t}e^{tx}dx$$

$$= \frac{2}{t}e^t - \frac{2}{t^2}[e^t - 1] = 2\frac{(t-1)e^t + 1}{t^2}$$

That this function is defined at zero can be checked with L'Hôpital's rule. That is, at zero it returns the value $\lim_{t \to 0} 2\frac{(t-1)e^t + 1}{t^2} = 2 \lim_{t \to 0} \frac{e^t + (t-1)e^t}{2t} = \lim_{t \to 0} e^t = 1$ as required (i.e. $r_0 = 1$).

The mean of a triangular distribution starting at zero, is 1/3 of *a* (the base length) back from its apex (by the rule for the centroid of a triangle). We wish now to perform a linear transformation upon *x* so that our distribution is flipped horizontally, rescaled and has a mean of one. This transformation can be written as ax + b, where a is the base length of our triangle ($0 > a \geq -3$) and b is a factor to fix the mean to one. Since the mean of this distribution is 1/3 |a|, then b = 1-1/3 a. Applying the well known result that if y = ax + b, then $M_y(t) = e^{bt}M_x(at)$, gives,

$$M_y(t) = e^{bt}\left[\frac{2}{at}e^{at} - \frac{2}{a^2t^2}[e^{at} - 1]\right]$$

$$= e^{t-at/3}\left[\frac{2}{at}e^{at} - \frac{2}{a^2t^2}[e^{at} - 1]\right].$$

Applying the same principle to translate a gamma distribution, let Y = zX + b (z > 1), where X follows a gamma distribution with mean fixed to one. The mean of zX is 1/z, so our translation is y = zx + 1-1/z (the last factor translating the mean is again one). The moment generating function of x is $((k - t) / k)^{-k}$, where *k* is the shape parameter. So the moment generating function of the translated gamma distribution with mean fixed to one is $M_y(t) = e^{bt}M_x(at) = e^{t-t/z}((k - zt) / k)^{-k}$, which setting z = 1 recovers the original moment generating function as expected. For any allowable value of *z* ($0 < z < 1$) the amount the original distribution is translated = 1 - 1/z.

Appendix 2.5 A closed form correction formula for a trimodal distribution

This is special case where we can infer a corrected distance for an underlying trimodal distribution of rates across sites, without needing to a priori separate sites into their correct rate classes. The p.d.f., $f_y(y)$, of the underlying distribution of sites has a proportion, p_0, of invariant sites (class 0), with the remaining sites falling into two rate classes (λ_1 and λ_2), such that the overall mean is one, and that $\lambda_2 = 2\lambda_1$. Therefore,

$$p_0 + p_1 + p_2 = 1 \text{ and } p_1\lambda_1 + p_2\lambda_2 = 1.$$

Let $z = e^{\lambda_1 t}$, so $M_y(t) = p_0 + p_1 z + p_2 z^2$ (in our case $M_y(t)$ is r_i). We can solve this equation to give a closed form inverse to the moment generating function, that is

$$p_2z^2 + p_1z + (p_0 - M) = 0,$$

$$\text{so } z = \frac{-p_1 \pm \sqrt{p_1^2 - 4p_2(p_0 - y)}}{2p_2},$$

$$\text{and } t = \frac{1}{\lambda_1} \ln\left[\frac{-p_1 + \sqrt{p_1^2 - 4p_2(p_0 - y)}}{2p_2}\right].$$

For our purpose here, we replace t with ρ_i and y with r_i, then chose values for parameters p_0 and p_1(note we have two parameters since $p_2 = 1 - p_0 - p_1$, and we require the restriction that all these values must be positive so $p_0 + p_1 \leq 1$). Having chosen values of p_0 and p_1 we then solve for λ_1,to fix the mean of the underlying distribution of rates across sites to 1. We require $p_1\lambda_1 + 2p_2\lambda_1 = 1$, so $p_1\lambda_1 + 2\lambda_1(1-p_0-p_1) = 1$,

$p_1\lambda_1+2\lambda_1-2p_0\lambda_1-2p_1\lambda_1 = 1$, so $\lambda_1(2-2p_0-p_1) = 1$, giving $\lambda_1 = 1/(2-2p_0-p_1)$.

Appendix 2.6 Order 2^{t-1} Hadamard conjugations for 4-state data

Early in the this PhD. project (1991) I developed a way to use order 2^{t-1} Hadamard conjugations with 4-state nucleotide data. This method gives exactly the same γ weights as the bipartition entries (which are also order 2^{t-1}) in the 4^{t-1} conjugations when the data is without sampling error and generated under the Kimura 3ST model or its submodels. These methods have been proven to be consistent in work with Mike Steel, and by Mike Hendy. In sampling situations and when the model is violated, the methods described here will give different results to the order 4^{t-1} conjugations. They appear to be more sensitive in visualizing violations of the model than the 4^{t-1} conjugations. In addition these order 2^{t-1} conjugations offer faster calculation and can be constrained to all the standard submodels of the generalised Kimura 3ST model, including a fixed transition / transversion ratio across the tree. They can incorporate a distribution of rates across sites in the same basic way as the 4^{t-1} conjugations.

The first step in generating these order 2^{t-1} γ vectors is to use the aligned 4-state sequences (without any deletions) to code for 3 separate **s** vectors (we call these $\mathbf{s}_{ac}$, $\mathbf{s}_{ag}$, $\mathbf{s}_{at}$). The first vector $\mathbf{s}_{ac}$ is formed by grouping states A and C (e.g. giving them state *x*), versus states G and T (giving them state *y*). This recoded data is then coded into an **s** vector with the same indexing and grouping of symmetric patterns (e.g. $s_1 = f\{xyyy\}/c + f\{yxxx\}/c = f\{xyyy + yxxx\}/c$) as with standard 2-state data. The same procedure applied to the 4-state data generates $\mathbf{s}_{ag}$ (after grouping states {AG} versus {CT}) and $\mathbf{s}_{at}$ (after grouping states {AT} versus {CG}). Each of these vectors are counting a set of implied changes in the site patterns, for example $\mathbf{s}_{ac}$ is counting implied substitutions A ↔ G, A ↔ T, C ↔ G, C ↔ T. The double ended arrow in A ↔ G indicates it could have been either A → G or G → A; these models do not differentiate. The term implied substitution here is used in a loose sense, with a refined definition given below.

Next each **s** vector is separately multiplied by **H** (with 2^{t-1} rows) to give an **r** vector. That is $\mathbf{r}_{ac} = \mathbf{H}\mathbf{s}_{ac}$, $\mathbf{r}_{ag} = \mathbf{H}\mathbf{s}_{ag}$, and $\mathbf{r}_{at} = \mathbf{H}\mathbf{s}_{at}$. These **r** vectors are counting implied substitutions on all sizes of pathsets from pairwise distances up (they contain these vectors in the form of {1 - twice the observed pathset count}). One linear operation on these **r** vectors generates a new vector

which counts just transitional changes on all pathsets, and so is analogous to the quantity "P" in Kimura's (1980) and (1981) equations. We will call this vector **a**:

$$\begin{aligned} a_i &= f[A\leftrightarrow G + C\leftrightarrow T]/c \text{ (observed proportion of changes on pathset } i) \\ &= 1/2\, f[(A\leftrightarrow G + A\leftrightarrow T + C\leftrightarrow G + C\leftrightarrow T) + (A\leftrightarrow C + A\leftrightarrow G + C\leftrightarrow T + G\leftrightarrow T) \\ &\quad - (A\leftrightarrow C + A\leftrightarrow T + C\leftrightarrow G + G\leftrightarrow T)]/c \\ &= 1/2\, f[\{AC\}/\{GT\} + \{AT\}/\{CG\} - \{AG\}/\{CT\}]/c \\ &= 1/2\, (([1 - r_{ac_i}]/2) + ([1 - r_{at_i}]/2) - ([1 - r_{ag_i}]/2)) \\ &= 1/2\, ([1 + 1 - 1 - r_{ac_i} - r_{at_i} + r_{ag_i}] / 2) \\ &= 1/4\, (1 + r_{ag_i} - r_{ac_i} - r_{at_i}), \end{aligned} \qquad \text{(A2.6-1)}$$

where "f "stands for frequency.Likewise we have,

$$b_i = A\leftrightarrow T + C\leftrightarrow G = 1/4\, (1 + r_{at_i} - r_{ac_i} - r_{ag_i}),$$

and

$$y_i = A\leftrightarrow C + G\leftrightarrow T = 1/4\, (1 + r_{ac_i} - r_{ag_i} - r_{at_i}).$$

When the index of r_i indicates it is measuring events on pairwise distance, then a_i, b_i and y_i are equivalent to Kimura's (1981) P, $\overline{Q}$ and R respectively. Here, they have been renamed since P, Q and R already have specific meanings in the Hadamard transform nomenclature of Penny and Hendy (1989).

A2.6.1 Corrected pathset lengths under different models

It is now possible to estimate the implied number (taking into account multiple changes) of transitions and transversions on each even sized path set (not just pairwise distances) using the following equations which are rearrangements of the equations of Kimura (1981) e.g. see Gojobori *et al.* (1990). For the generalised Kimura 3ST model, the expected number of transitions on each pathset is,

$$\alpha = -1/4(\ln[1-2\mathbf{a}-2\mathbf{b}] + \ln[1-2\mathbf{a}-2\mathbf{y}] - \ln[1-2\mathbf{b}-2\mathbf{y}]), \qquad \text{(A2.6.1-1)}$$

where the natural logarithm (ln) is applied to each component in turn. Similarly, the number of transversions type 1 on each pathset is given in,

$$\beta = -1/4(\ln[1-2\mathbf{a}-2\mathbf{b}] - \ln[1-2\mathbf{a}-2\mathbf{y}] + \ln[1-2\mathbf{b}-2\mathbf{y}]), \qquad \text{(A2.6.1-2)}$$

and the number of type 2 transversions on each pathset is given by,

$$\psi = -1/4(-\ln[1-2\mathbf{a}-2\mathbf{b}] + \ln[1-2\mathbf{a}-2\mathbf{y}] + \ln[1-2\mathbf{b}-2\mathbf{y}]), \qquad \text{(A2.6.1-3)}$$

(where in all cases the logarithmic function is applied componentwise). In the case of pairwise distances Kimura (1981) labeled the quantity ψ as γ; here it is renamed ψ to avoid confusion with the γ vectors associated with Hadamard conjugations. Note that Kimura (1981) refers to our α_i as $2\alpha_i$, our β_i as $2\beta_i$ etc. as he considers them to be rate matrix entries estimated going back to a last common ancestor, assuming a clock.

These equations are equivalent to those in Kimura (1981), but have not previously been applied in such a direct manner to estimate the lengths of non-intersecting paths between 4 or more taxa. Notice the form of each of the 3 components (e.g. ln[1-2**a** -2**b**]) that are transformed then rearranged to give α. There is an equivalence between the **r** vectors and these terms, that is $\mathbf{r}_{ac}$ = [1-2**a**-2**b**], $\mathbf{r}_{at}$ = [1-2**a**-2**y**], and $\mathbf{r}_{ag}$ = [1-2**b**-2**y**]. This leads to a computational short cut in a specific case which is described later.

If the constraint **b** = **y** is applied, this forces the corrections to Kimura's (1980) 2ST model (in generalised form) and the corresponding formulae are:

$$\alpha_{K2} = -1/2 \ln[1-2\mathbf{a}-(\mathbf{b}+\mathbf{y})] + 1/4 \ln[1-2(\mathbf{b}+\mathbf{y})] \quad \text{(A2.6.1-4)}$$

$$(\beta + \psi) = -1/2 \ln[1-2(\mathbf{b}+\mathbf{y})]. \quad \text{(A2.6.1-5)}$$

Note that in this section we use the brackets ( ) around the quantities α, β, and ψ to indicate a single joint estimate is being made. Here α_{K2} estimates the transitions on each pathset, while (β + ψ) is the Kimura 2-parameter estimate of the number of transversions. The expected number of transversions estimated this way is identical to the 2-state Hadamard conjugation (which uses the Poisson or Cavender model) when the data is first coded into R vs Y (to be precise (β_i + ψ_i) = -1/2 (R/Y 2-state ρ_i)). Similarly to the generalised Kimura 3ST model, the relative rate of transitions to the single rate for transversions can vary across the tree, and these order 2^{t-1} Kimura 2ST Hadamard conjugations will still be consistent as long as each edge has a transition matrix of the Kimura 2ST form (this goes for pairwise distance estimates also, i.e. they are still additive).

If we constrain **a** = **b** = **y**, then the implied length of each pathset (transitions and transversions not distinguished) is estimated using the Jukes-Cantor equation (Jukes and Cantor 1969, Kimura and Ohta 1972) which is a 1-parameter model (the possible advantages of fewer parameter models are discussed later in the thesis). That is,

$$(\alpha+\beta+\psi) = -3/4 \ln[1- 4/3 (\mathbf{a} + \mathbf{b} + \mathbf{y})], \quad \text{(A2.6.1-6)}$$

where the logarithmic function is again applied componentwise (as in all other equations in this appendix).

A number of other restrictions and / or generalizations can easily be made to the pathset length corrections. Most of the family of corrections associated with Kimura's 3 ST model can be used, and the results will be exact under that model. To allow for a distribution of rates across sites, one simply replaces the *ln* (natural logarithm) function with M^{-1} (the inverse of the moment generating function of the distribution of site rates, more on this in the case of pairwise distances in chapter 3). Golding (1983) and Jin and Nei (1990) give examples of such transformations for the Jukes-Cantor and generalised Kimura 2ST models when there is assumed to be a Γ distribution of rates across sites. It is also possible to force corrections to the "homogeneous" Kimura 2ST or 3ST models. This is appropriate when there is no evidence to reject the ratio of transitions to transversions being fixed on all edges of the tree. An appropriate way to infer these "homogeneous" distance estimates is by ML as described and used in PHYLIP (Felsenstein

1993). For each distance it involves predicting the observed proportions of changes of the various types (e.g a and $(b + y)$ for the 2ST model) for a specific distance given a ratio of α to $(\beta + \psi)$, then finding the model distance that best predicts the proportions of a and $(b + y)$ (in place of ML, a Pearson X^2 measure could be used to measure the fit of observed to predicted a and $(b + y)$). It may be convenient to store the predicted proportions of a and $(b + y)$ in a look up table, or as a spline function (a mixture of lower order polynomial approximations) if many pathsets must be corrected (e.g. there are up to $\approx 3 \times 10^6$ with 21 taxa).

Other pathlength corrections which are not exact under a model may also be used. Schöniger and von Haeseler (1993), for example, consider weighting transversions more heavily then transitions (however as is apparent later in the thesis, it is probably statistically more efficient to do the transformations, then weight the entries in the γ vectors by the inverse of their standard deviations). Another appealing extension would be to distinguish further types of changes and apply corrections which are exact under more than 3 parameter models. This approach will be approximate, since Székely *et al.* (1993) show that in order to exactly infer the length of pathsets with more than two endpoints (without knowing the tree), then with 4-state data the model must be the generalised Kimura 3ST or a submodel. This approach is not pursued here, but it may yield useful approximations in some circumstances. The calculations of Penny *et al.* (1990) are of this nature.

A2.6.2 Multiplication of corrected pathset length vectors to obtain γ vectors

To convert implied pathset lengths into model "corrected" bipartitions (called gamma vectors) the inverse Hadamard transform ($\mathbf{H}^{-1}$) is used. For the 3 parameter model multiply -2α by $\mathbf{H}^{-1}$ to give γ_α. Do likewise for β and ψ to give $\gamma\beta$ and γ_ψ. These 3 gamma vectors respectively estimate the number of transitions and transversions type 1 and type 2 that occurred on each of the 2^{t-1} bipartitions that can be used for tree selection. These bipartitions are also invariants since, under the model, if they are not compatible with the tree that generated the data, then they go to zero as the sequences become longer. Likewise the γ vectors for the 1 and 2 parameter models contain families of invariants. When converting corrected pathset lengths into g using the inverse Hadamard transform, $\mathbf{H}^{-1}$, first multiply by -2 to obtain γ of the correct magnitude and sign, e.g. for the Jukes Cantor $\gamma = \mathbf{H}^{-1}(-2(-3/4\ln[1-4/3(\mathbf{a}+\mathbf{b}+\mathbf{y})])) = \mathbf{H}^{-1}(3/2\ln[1-4/3(\mathbf{a}+\mathbf{b}+\mathbf{y})]))$, where the log function is applied componentwise.

There is also a computational short cut if desired corrections are limited to the generalised Kimura 3ST model. The usual form of the Kimura 3ST distance correction is $\alpha_i + \beta_i + \psi_i = -1/4\ln[(1-2a_i-2b_i)(1-2a_i-2y_i)(1-2b_i-2y_i)]$, where $(\alpha + \beta + \psi)_i$ is designated K, in equation 21 of Gojobori *et al.* 1990). On any pathset this can be expressed as

$$\begin{aligned}\alpha_i + \beta_i + \psi_i &= -1/4\ln[(r_{ac})_i \times (r_{at})_i + \times (r_{ag})_i] \\ &= -1/4\{\ln[(r_{ac})_i] + \ln[(r_{at})_i] + \ln[(r_{ag})_i]\}\end{aligned} \qquad \text{(A2.6.2-1)}$$

So we may form ρ_{ac}, ρ_{at}, ρ_{ag} by taking the natural log componentwise of $\mathbf{r}_{ac}$, $\mathbf{r}_{at}$, $\mathbf{r}_{ag}$, respectively, then estimate α, β, and ψ at the level of ρ. Thus ρ_{ac} counts -2 the corrected number of transitions and transversion type 1 on each pathset, so $\rho_{ac} = -2(\alpha + \beta)$, and similarly $\rho_{at} = -2(\alpha + \psi)$, $\rho_{ag} = -2(\beta + \psi)$, so $\alpha = -1/2(\rho_{ac} + \rho_{at} - \rho_{ag})$, and similarly for β and ψ. Even more simply we can multiply each of these ρ vectors separately by $\mathbf{H}$, then rearrange these vectors i.e. $\gamma_{ac} = \mathbf{H}^{-1}\ln(\mathbf{Hs}_{ac})$, $\gamma_{ag} = \mathbf{H}^{-1}\ln(\mathbf{Hs}_{ag})$, and $\gamma_{at} = \mathbf{H}^{-1}\ln(\mathbf{Hs}_{at})$, while $\gamma_{\alpha} = \gamma_{ac} + \gamma_{at} - \gamma_{ag}$, $\gamma_{\beta} = \gamma_{ac} + \gamma_{ag} - \gamma_{at}$, and $\gamma_{\psi} = \gamma_{ag} + \gamma_{at} - \gamma_{ac}$. Here again the natural logarithm can be replaced by the inverse of the moment generating function of the distribution of rates across sites (M^{-1}), which has been standardised and scaled as in table 2.2.

An interesting point is that $\gamma_{ag} = \mathbf{H}^{-1}\ln(\mathbf{Hs}_{ag})$ is equal to γ from the 2-state Hadamard conjugation applied to the recoding R / Y (which is just grouping {AG} versus {CT}). Earlier we noted that the count of transversions $(\beta + \psi)_i$ under the Kimura 2ST model was also equal to the 2-state conjugation after grouping R / Y. This in turn implies that $\gamma_{(\beta + \psi)} = \gamma_{(R\,/\,Y)} = \gamma_{ag}$, that is the sum of both types of transversion estimated under the Kimura 3ST model is the same as the number of transversions estimated under the Kimura 2ST model. This can be shown algebraically,

$$\begin{aligned}
\beta_i + \psi_i = \quad & -1/4(\ln[1-2\mathbf{a}-2\mathbf{b}] - \ln[1-2\mathbf{a}-2\mathbf{y}] + \ln[1-2\mathbf{b}-2\mathbf{y}]) && \text{(from equations A2.6.1-2 and -3)} \\
& + -1/4(-\ln[1-2\mathbf{a}-2\mathbf{b}] + \ln[1-2\mathbf{a}-2\mathbf{y}] + \ln[1-2\mathbf{b}-2\mathbf{y}]) \\
= \quad & -1/4(\ln[1-2\mathbf{b}-2\mathbf{y}] + \ln[1-2\mathbf{b}-2\mathbf{y}]) \\
= \quad & -1/2(\ln[1-2\mathbf{b}-2\mathbf{y}]) \\
= \quad & -1/2(\ln[1-2(\mathbf{b}+\mathbf{y})]) && \text{(from equation A2.6.1-5)} \\
= \quad & (\beta + \psi)_i
\end{aligned}$$

That the estimated total number of transversions is equal under both the generalised Kimura 2ST and 3ST models is a result of states being groupable for certain Markov models, which is dealt with in more detail in section 3.7.7. In contrast, the estimates of the number of transitions under these two models is different, i.e. $\alpha_{K3} \neq \alpha_{K2}$ since equation A2.6.1-1 ≠ A2.6.1-4 (under the more general models A2.6.1-1 is typically greater than A2.6.1-4 since when b ≠ y, -1/4(ln[1-2a-2b] + ln[1-2a-2y]) will be greater than -1/2ln[1-2a-b-y] due to the convex nature of the transform).

A2.6.3 Counting changes on higher order pathsets, and proving consistency

It is possible to understand the intermediate steps in these 4-state Hadamard conjugations in a similar way to the correcting of observed pairwise patterns that Kimura (1981) considered. If you look at what the elements in **a**, **b** or **y** are on higher order pathsets, then a very simple pattern emerges. Looking at the nucleotide pattern for a single site on an even sized set of taxa, then the rules are

(a) Remove in pairs states A, C, G and T (repeat until there is either 1 or 0 of each state left).

(b) If only two unpaired states remain, then this records an event of one of type *a*, *b*, or *y*.

(c) If either four unpaired states or no states are left, then this counts as "record no event."

For example, if the tips of a pathset have the states CGACTTCG, remove all the pairs of states (here they are a pair of C's, a pair of G's and a pair of T's), and you are left with AC. So this AC residual unambiguously indicates an "observed" change A↔C (we say unambiguous because unless we know the tree, then we cannot exclude that all the pairs of states are adjacent to each other on the tree, and are thus implying no changes on that part of the path). This A↔C change adds 1/c to the size of the corresponding entry in *y* for this pathset (i.e. a type 2 transversion). If a site showed the following pattern at the tips of a pathset ACGTAA, then after removing the pair of A's, we are left with CGTA, or four unpaired states. In this case this site does not contribute to increase the size of this pathsets entry in either *a*, *b*, or *y*. This is because it is not possible, given only this information and not the tree, to identify which changes are indicated.

All these new conjugations claimed to be exact under the model have been checked numerically with double precision calculations, by comparing the quantities they estimate with their counter parts in the full 4^{t-1} 4-state Hadamard conjugation. In addition mathematical proofs within the context of order 4^{t-1} 4-state Hadamard conjugations have been constructed in collaboration with Drs Mike Steel and Mike Hendy. These will be given elsewhere. It would be interesting to derive equivalent proofs in the same manner that Kimura used to prove his original corrections on pairwise distances.

A2.6.4 Applications to data

These order 2^{t-1} conjugations allow great versatility in the ways sampling variance might be reduced from the order 4^{t-1} conjugations. They allow the model to be forced to specific submodels, and they may also decrease sampling variance by counting and relating each type of implied substitution to just the weights of possible edges in the tree (the bipartition entries in the 4-state Hadamard conjugation).

Applying these conjugations to the data of Horai *et al.* 1992 has been particularly instructive. The order 2^{t-1} spectra for transitions and types 1 and 2 transversions look similar to those in figure 2.9. However because all the changes in the off diagonal elements are also focused into just bipartition entries, there is more obvious evidence of contradictions to the human-chimp signal under i.r. models. For example a pattern like ACAT for human, chimp, gorilla and orangutan respectively, is largely hidden from the gaze in the order 4^{t-1} conjugation as a relatively rare model invariant. However under the 2^{t-1} it implies a transversional change separating human and gorilla from chimp plus orangutan. The same applies with the transitions. Consequently, with the Horai *et al.* (1992) data the non-tree transition bipartition signals are generally twice as large as calculated with the 4^{t-1} conjugation (as shown in figure 2.9), yet the signals compatible with the human-chimp tree remain of a similar size (order 2^{t-1} results not shown). When corrections assuming a Γ distribution of rates across sites are then applied (e.g. with $k = 0.35$), the change is more dramatic than seen with the 4^{t-1} conjugation, and the bipartitions clearly become much more tree like, in this case again supporting the human-chimp tree.

These conjugations can also be run in reverse to predict the vectors (T) $\mathbf{s}(T)_{ac}$, $\mathbf{s}(T)_{ag}$, $\mathbf{s}(T)_{at}$ for a weighted tree under any submodel of the generalised Kimura 3ST model with unequal rates across sites. Unfortunately as yet we have not found a way to predict all the 4^{t-1} entries in $\mathbf{s}(T)$ from the values in just the order 2^{t-1} $\mathbf{s}(T)$ vectors. This precludes making the normal type of ML calculations. A variant form of ML can however be run, by summing the fit of each observed and expected $\mathbf{s}$ vector to give an overall fit. We do not claim this variant of ML to be exact in any specific way (although it could be, as this has yet to be checked). It does however allow direct weighting of the transversional changes, over the transitional changes (the vector $\mathbf{s}_{ag}$ for example counts all observed transversional difference in the data). This could be useful when transitions are near saturation, or experiencing more systematic error than the transversional changes (or perhaps the reverse).

Lastly, the variance covariance matrix of each $\mathbf{r}$ vector can be obtained by standard methods, like those used in Hasegawa *et al.* (1985), Bulmer (1991) and Waddell *et al.* (1994). These estimate the variances and covariances between pathset lengths in accordance with the proportion of varied sites which simultaneously imply changes on two pathsets. Next, delta method approximations like those used for the standard variance estimates of the Kimura distances (e.g. see Gojobori et al. 1990) are made to estimate the variance-covariance matrix of the ρ vectors. This is followed in turn by more linear operations to estimate the covariance matrix of all entries in the separate γ vectors. At present we have not found a neat computational form for the covariance matrix of all entries in these separate γ vectors, such as with the standard Hadamard conjugation (for more detail see chapter 4), although the necessary calculations can certainly be made using standard algorithms with today's computers. The bootstrap offers another way to make these estimates (although biased towards overestimating model variances and covariances due to the non-linear transforms). Overall, we anticipate these new conjugations will offer biologists a variety of useful ways to view their data after making "corrections" but before tree selection.

CHAPTER 3:

MODIFYING THE LOGDET DISTANCE TO COPE WITH UNEQUAL RATES ACROSS SITES

"At best, we must discount the precision of our estimates of the phylogeny according to our skepticism of the details of the model of evolution,"

[Felsenstein 1981b]

3.1 INTRODUCTION

This chapter focuses on the exciting new development of log determinant (LogDet) distance measures. Here we review their history, describe them in the context of Markov models, show how they may be made robust to unequal rates at different sites, look briefly at their statistical properties, and then use them to analyse sequences spanning the "tree of life."

There exist many i.i.d. Markov models for which a Hadamard conjugation may not reliably remove the effect of multiple hits, as they are limited to making exact corrections when transition matrices on each edge in the tree have an Abelian group structure (e.g. the generalisied Kimura 3ST model). In reality it is expected that the probabilities of all twelve types of substitution (A $\rightarrow$ C etc.) will vary through time, and independently in different lineages, as the cellular environment and particularly the DNA replication / repair mechanisms evolve. This not only makes distances hard to estimate, but can also dramatically distort their relative sizes either with or without one of the simple distance transformations presently used (for example the Poisson distance correction of Jukes and Cantor 1969, or the two parameter method of Kimura 1980). A tree reconstruction algorithm such as neighbor joining can then be mislead by such data and select an incorrect tree. One indicator that the relative substitution probabilities have changed within a set of sequences is that they have significantly different base compositions. This symptom of varying base composition, can become an immediate cause of errors with standard distance methods (e.g. Saccone *et al.* 1989, Weisburg *et al.* 1989, Lockhart 1990, Loomis and Smith 1990, and especially Lockhart *et al.* 1992 and 1994), because they assume stationarity of base composition and / or and a specific subset of possible rate matrices (especially those giving rise to a time reversible models).

One of the most exciting developments in the past two years has been recognition that by taking the logarithm of the determinant of a matrix of the frequencies of nucleotide pairs in sequence i and sequence j gives a tree additive distance under the most general (mild) set of assumptions so far (Steel 1994a, Lockhart *et al.* 1994, and Lake 1994 (who calls the method "paralinear distances")). Basically this "LogDet" distance measure, combined with a reliable tree estimation algorithm, will be consistent for any i.r. and i.i.d. tree model of sequence evolution

(where i.r. is identical rates and i.i.d. states sites are independent and identically distributed, as defined in chapter 2). This model is mathematically defined in Steel (1994a); it is a very general i.i.d. model that allows base composition to vary from species to species.

A major drawback of the LogDet method being applied to coding sequence is these transformations are only additive when all sites evolve at an identical rate. Consequently LogDet distances are as sensitive to unequal rates across sites as other i.r. distance measures as we will show in this chapter (and contrary to the claims of Lake 1994 in this respect). As we have already demonstrated in chapter 2, violation of the i.r. assumption is a major cause of inconsistency in tree estimation, especially with (but not limited to) coding sequences. In this chapter we look further at ways in which we can approximate the effect of a continuous distribution of rates across sites by removing an optimal proportion of constant sites (whether they are truly invariant or just slowly changing). Combining the LogDet transformation with an invariant sites model appears to be a reliable way of obtaining a robust distance estimate when rates vary across sites (an alternative way which we also discuss is to separate sites into rate classes). We show that application of "invariant sites-LogDet" to Gouy and Li's (1989) data set of 16S-like rRNA leads to important new insights on the evolution of ancient lineages in the "tree of life." This section also involves application of all the described methods of inferring what proportion of sites should be treated as invariant (that is unable to change at all), and also includes the development of a new "capture-recapture" statistic to do this.

Another question also looms before we can confidently use the LogDet with real data, and this is the question of what is its sampling error like in realistic applications (sampling error is the result of sampling variance plus bias). Being a highly general distance measure it may have the same failing that appears to plague all distance measures as we add further parameters to the underlying model; that is it has variances which become large enough to seriously reduce the probability of reconstructing the correct tree (e.g. the results of Rodríguez *et al.* 1990, considered in the light of Hillis *et al.* 1994). A direct resampling approach (the bootstrap) is used to do this.

Lastly the history of LogDet type methods applied to sequence data is surprisingly large, and stretches back to before 1987. We list the achievements of these researchers as they to are relevant to the derivation of our current understanding of this class of methods, and should be recognised for their insights.

Here now is a brief overview of the structure of this chapter. After an introduction to Markov processes on trees, we discuss the history of LogDet methods in phylogenetics, and show that despite useful work some of the earlier claims were inaccurate. We give an improved interpretation of what LogDet distances mean and show how we can interpret these values biologically. We then consider different ways we can remove invariant sites, so as to make an i.r. distance transformation more robust to unequal rates of change at different sites. We conduct a phylogenetic analysis of the 28 taxa 16S-like rRNA data set of Gouy and Li (1989), which remains very much at the center of research into the so called "tree of life." In particular we compare our statistical estimate of support for the monophyly of the archaebacteria with that of Gouy and Li (1989). We study this data set in detail not only to learn about the evolution of

rRNA molecules, but also about how the modified "invariant sites LogDet" transform performs on such data. Illustrated are ways of estimating the distribution and variability of base compositions, estimating the proportion of constant sites that should be removed in order to make the data best fit to the i.r. expectations, and an assessment of how large the sampling variances and bias in the LogDet transform are compared to the Jukes-Cantor distance correction.

To improve the reliability of our statistics, we set up six hypotheses about the "tree of life" prior to analysis. We then compare the results we obtain with invariant sites and LogDet models, with those of an otherwise identical analysis using the simple, but very popular identical rates Jukes-Cantor distance. We find one especially surprising result which argues quite strongly that Microsporidia, and not diplomonads (such as *Giardia*), are the most ancient branching in the eukaryotic kingdom. Overall the results are very encouraging and suggest that "invariant sites LogDet", combined with a reliable tree building algorithm, may offer the most robust method for estimating a tree of relationships from highly diverged (or ancient) functional molecules.

This a list of some specific terms used in this chapter.

Symbol	Definition
$\mathbf{F}$	A divergence matrix, which counts the frequency of all dinucleotide patterns (i.e. AA, AC, ... TT) between two aligned sequences
$\mathbf{F}^*$	An $\mathbf{F}$ matrix which has been symmetrised by averaging entries ij and ji
$\hat{\mathbf{F}}$	An $\mathbf{F}$ matrix estimated from a sample
$\mathbf{F}^{\#}$	An $\mathbf{F}$ matrix where entries expected to be equal under the Kimura 3ST model have been averaged (e.g. transitions all equal, AG = GA = CT = TC).
$\mathbf{P}$	A matrix of transition probabilities
$\mathbf{R}$	A transition rate matrix
Π	A diagonal matrix of normalised (i.e. sum to one) nucleotide base frequencies

3.2 FUNDAMENTAL EQUATIONS OF A MARKOV PROCESS ON A TREE

This section introduces many of the standard probabilistic equations that underpin our i.i.d. models of sequence evolution, before moving on to describe new results beginning in section 3.2.2. The results in this introduction come from standard Markov theory. The mathematical statistical background to these equations is standard, and can be found in books such as Bellman (1960), Karlin and Taylor (1975), Keilson (1979), or Iosifescu (1980).

The primary result of sequence evolution is that after a time, some sites are observed in a different state to their ancestral one. Thus the most fundamental equation for a Markov process on an edge going from node m (nearest the root) to node n is $\mathbf{F}(mn) = \Pi(m)\mathbf{P}(mn)$, where $\mathbf{F}$ is the divergence matrix, and it contains the probabilities (or normalised frequencies) of state i at point m being in state j at point n, $\Pi(m)$ is a diagonal matrix of the normalised (so they sum to one) frequencies of all states at point m, while $\mathbf{P}(mn)$ is a matrix of the conditional probabilities

of state *i* changing to state *j* between points *m* and *n* (by conditional we mean the probability of change of state given that *i* is already in that state, so consequently all rows in **P** sum to one). The term conditional is often dropped in this context, and **P** is typically called the transition probability matrix. This equation is applicable even when the frequencies of the different states are not in equilibrium at point *m* or *n*. If the process of evolution is Markovian (i.e. random or stochastic) then the determinant of **P** will always be greater than zero, and tending to zero as the amount of diverge (number of substitutions) goes to infinity. If we have a series of edges between nodes *m* and *n* moving away from the root towards a tip in the tree (say edges 1, 2 .. ,x), then $\mathbf{F}(mn) = \Pi(m)\mathbf{P}(1)\mathbf{P}(2)...\mathbf{P}(x)$. In the most general i.i.d. model the multiplications are non-commutative. An example of this equation is shown in figure 3.1, equation (1).

We now consider how we get a matrix **P**. If we have a continuous time process (i.e. changes can be thought of as happening in very fine intervals of time) then **P** represents exp(**R**t), where exp is the matrix exponent operation, **R** is the instantaneous rate matrix operating between points *m* and *n* (that is a matrix of the relative rates at which substitutions are occurring), while *t* is a scalar (not necessarily linear with time) controlling how much divergence has occurred (given that **R** is set first). Because **R** measures rates of change, all off-diagonal elements in **R** must be ≥ 0, and each row sums to zero. The definition of a matrix exponent is $\exp(\mathbf{R}t) = \sum_{n=0}^{\infty} \frac{(Rt)^n}{n!} = \mathbf{I} + (\mathbf{R}t)/1! + (\mathbf{R}t)^2/2! + , \ldots , + (\mathbf{R}t)^n/n! + \ldots$ (if this convergences). The matrix exponent is calculating the net effect of process **R** at each instant of time, as it accumulates over a certain period (not real time, best thought of as intervals of change). Treating time as being fine grained (for example 1000 equal spaced intervals of change), then,

$$\exp(\mathbf{R}t) \approx (\mathbf{I} + \mathbf{R}t/1000)^{1000} = (\mathbf{I} + \mathbf{R}t/1000)_1 \times (\mathbf{I} + \mathbf{R}t/1000)_2 \times ..... \times (\mathbf{I} + \mathbf{R}t/1000)_{1000},$$

where **I** is the identity matrix (1's on the diagonal, 0 elsewhere). Every member of the series (**I** + **R**t/1000) is the transition matrix in a short interval of time (t/1000). This relationship becomes exact as **R**t is divided up an infinite number of times. Conveniently, if (**R**t) is able to be diagonalised it can be calculated as $\exp(\mathbf{R}t) = \Omega\exp(\Psi)\Omega^{-1}$, where Ω is a matrix of the right eigenvectors of (**R**t), Ω^{-1} is its inverse (which here is the transpose, as eigenvectors are orthogonal), Ψ is a diagonal matrix of the eigenvalues of (**R**t), and the exponent function is applied to each diagonal element of Ψ in turn (i.e. componentwise).

In the case of 2-state characters, all **P** matrices with positive determinant can be expressed as exp(**R**t) for some **R** (see appendix 3.1 for a proof). With three or more states, **P** matrices which correspond to a homogeneous continuous time process (i.e. just one) **R** matrix, are a strict subset of all possible **P** matrices. Another useful matrix algebra results for programming these stochastic processes are that any product **R**t can be diagonalised using the eigenvector approach (as long as it does not have degenerate eigenvalues, i.e. two identical eigenvalues). Further this diagonalisation always gives eigenvalues such that the largest is 1, while the remaining ones are all negative. Some of these negative values may be complex numbers, but their exponents may

be calculated using trigonometric functions and come out to be real numbers between zero and one.

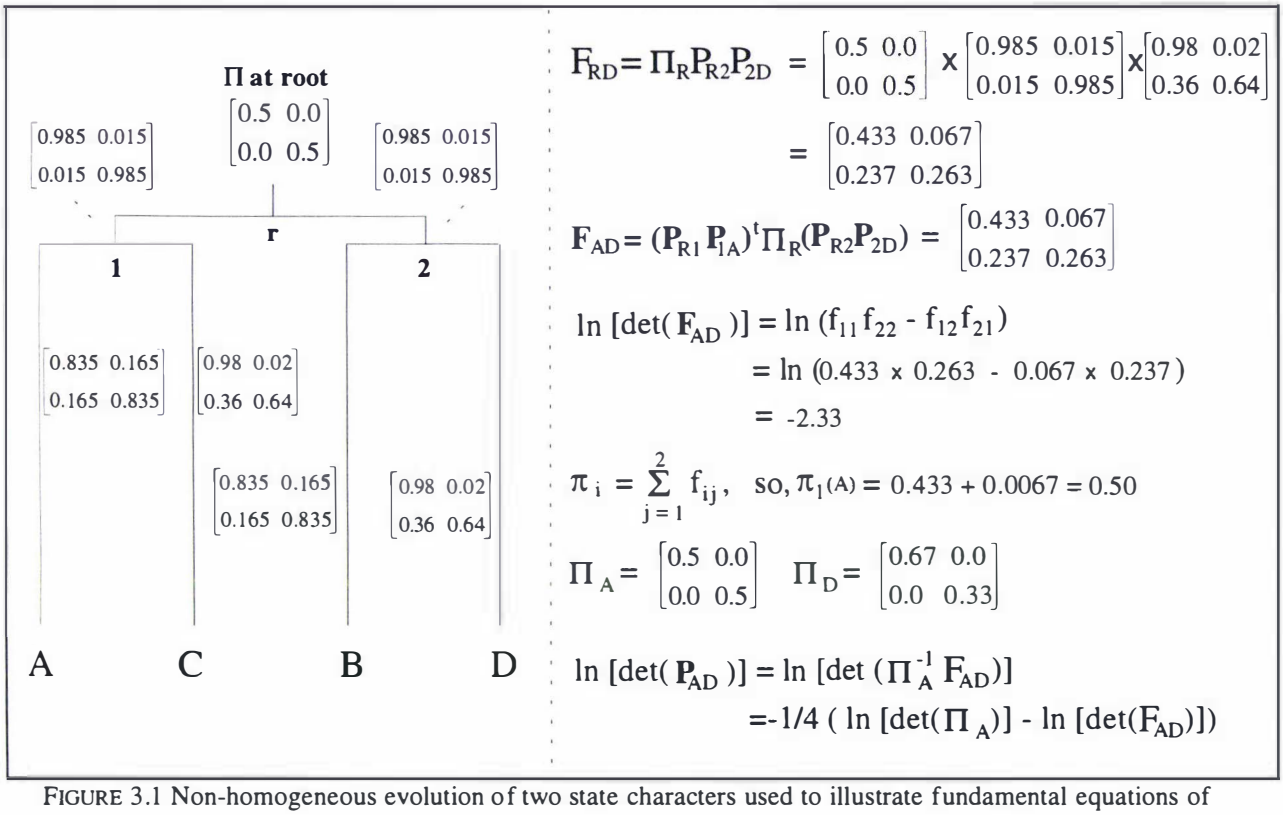


FIGURE 3.1 Non-homogeneous evolution of two state characters used to illustrate fundamental equations of a Markov model. The tree of 4 taxa (A, B, C, D) has equal internal edge lengths, while the external edges are also of equal length (by equation 3.2-1). Associated with each edge of the tree is a transition matrix (P) of the probabilities of seeing a substitution comparing a sequence going from the node at the top (e.g. node 1) to the bottom of an edge (e.g. node C). In the text we call this model a "quasi-clock", because while the number of substitutions is equal amongst all lineages, the rates of evolution amongst states are not (this model is described in more detail in section 3.4.1). Next to the tree are examples of some fundamental Markov equations on this tree.

It has been shown that any **R** matrix which defines a time reversible process, has all real eigenvalues (e.g. Keilson 1979). This follows since any time reversible process can be written as the product of a diagonal matrix Π, and a symmetric rate matrix, **S**, i.e. **R** = Π**S**, (see Tavaré 1986, with a proof of this in Waddell and Steel unpublished)(where Π also turns out to be the vector of stationary base frequencies). Given that both Π and **S** are symmetric, it follows from a simple result (see for example section 3.2, p.32 of Keilson 1979) that this matrix always has real eigenvalues. As we will see in the next section, the product of a time reversible process, the **F** matrix, is also symmetric so it to has real eigen values. Knowing this helps us by being able to use more specific (quicker and reliable) numerical methods to find the eigenvectors and eigenvalues of such matrices. Keilson (1979, section 3.2, p. 34) also shows that any time reversible process has complete monotonicity; in our context this means that as the true distance increases, so to must the observed distance (and vice versa).

Given a stationary continuous-time process, and knowing Π, **R** and t (here time), how does one estimate the number of substitutions of each type that have occurred on that edge? (Note that

stationarity implies that $\pi\mathbf{R} = 0$, where π is a row vector of proportions of character states, and 0 is a row vector of all 0's). If the base composition is in equilibrium the solution is simple, the proportion of *i* to *j* changes per site is just $\pi_i r_{ij} t$. Since each diagonal element of **R** is minus the sum of its row, then the total number of substitutions (often noted as δ_{ij}) from node *m* to *n* is just -tr($\Pi\mathbf{R}t$) (where tr is the matrix trace, the sum of the diagonal elements). This measure is additive and so can be can be summed over edges in the tree e.g. if edge e(m,n) (with length t_1) then joins to edge e(n,o) (length t_2), then the total number of changes is -tr($\Pi\mathbf{R}t_1$) + -tr($\Pi\mathbf{R}t_2$) = -tr($\Pi\mathbf{R}[t_1+t_2]$). If Π is not in equilibrium with **R** the solution of δ_{ij} involves an integral through time, as Π changes. For just two states (x and y) the total number of substitutions on an edge can be found explicitly as

$$\delta_e = 1/p(p^2_{xy}\pi_x + p^2_{yx}\pi^2_y - p_{xy}p_{yx} - [2p_{xy}p_{yx} \times \ln(1-p) / p]) \qquad (3.2\text{-}1)$$

where $p = (p_{xy} + p_{yx})$, and p_{ij} is the conditional probability of seeing state *i* replaced by state *j*, and π_i is the starting frequency of state *i* (Steel *et al.* 1993a).

Consider now estimating **F**, the divergence matrix between sequences *m* and *n*, under general i.r. / i.i.d. conditions. Suppose sequences *m* and *n* are at the tips of a rooted tree, and where *r* is the root of this path (i.e. last common ancestor of *o* and *p*), then $\mathbf{F} = \mathbf{P}^t(rm)\Pi(r)\mathbf{P}(rn)$, where $\mathbf{P}^t(rm)$ is **P**(rm) transposed. If **P**(rm) is made up of a series of transition matrices (a, b, .. n), then $\mathbf{P}^t(rm) = [\mathbf{P}(a)\mathbf{P}(b)...\mathbf{P}(n)]^t = (\mathbf{P}^t(n)...\mathbf{P}^t(b)\mathbf{P}^t(a))$ a standard matrix algebra result. An example of such a calculation in figure 3.1. The actual summation of probabilities occurring via this matrix multiplication is $f_{ij} = \sum\pi(r)_x p(ri)_{ix} p(rj)_{xj}$, where the summation is over *x* all possible nucleotide states (i.e. A, C, G, T) at the root of the path (r). This summation is the same calculation that occurs with maximum likelihood estimation of sequence patterns, except that when inferring sequence pattern probabilities amongst more than two sequences, the summation is over all internal nodes (not just the root) (Felsenstein 1981a).

3.2.1 A general distance estimate for time reversible models

Here we review distances under the general time reversible model before extending them to allow for a distribution of rates across sites. It is possible to solve for δ, the number of substitutions per site between sequence *m* and *n* (= -trace(**R**t)), given just **F** only if we assume some constraints. Here is one approach to making these constraints. We need to be able to rearrange the equation so $\mathbf{F} = \Pi(r)\mathbf{P}(mr)\mathbf{P}(rn)$, implying that $\mathbf{P}^t(rm)\Pi(r) = \Pi(r)\mathbf{P}(mr)$. This implies time reversibility of the process (i.e. the result is identical going either up or down each edge) and consequently **R** is of a special form (and can have a maximum of nine free parameters, or six if we specify π, e.g. Tavaré 1986). Next we require,

$\mathbf{P}(rm)\mathbf{P}(rj) = \exp(\mathbf{R}t_1)\exp(\mathbf{R}t_2) = \exp(\mathbf{R}(t_1+t_2))$, where t_1 is the scalar for evolution from r to *m*, and t_2 is the scalar for evolution from r to *n*. This requires that **R** is the same in both lineages (with two special exceptions mentioned below). And lastly, it is necessary assume that $\pi(r)$ is at

the stationary value for the matrix **R**. Recounting that $\mathbf{F} = \Pi\mathbf{P}$, (where we estimate Π from the observed nucleotide frequencies in **F**), implies that $\mathbf{P} = \Pi^{-1}\mathbf{F}$. Recall also that $\ln(\mathbf{P}) = \mathbf{R}(t_1 + t_2)$ and δ = -trace $\pi(\mathbf{R}t)$. Putting all this together we have

$$\delta = -\text{trace}[\Pi \ln(\Pi^{-1}\mathbf{F})] \qquad (3.2.1\text{-}1)$$

Using just this reasoning no fewer than 4 sets of authors have independently presented this solution (in varying degrees of compactness) as an original result (Taveré 1986, Barry and Hartigan 1987b, Rodríguez *et al.* 1990, and in a slightly different but equivalent form by Lanave *et al.* 1984).

The first paper to give a formula to solve for the general time reversible model was Lanave *et al.* (1984), although they incorrectly attributed it as a solution to any homogeneous 12 parameter model of sequence evolution (with stationary base composition). Rodríguez *et al.* (1990) gave equation 3.2.1-1 as a solution to any time reversible model, but added the unnecessary assumption a molecular clock. They did point out that there is one other set of models under which this method solves for δ, namely any equi-frequency model (i.e. π at equilibrium has elements 1/x, where x is the number of states). Also under equi-frequency models, **R** can be conditionally non-homogeneous (specifically each time **R** changes it must be directly to another equi-frequency rate matrix), and the formula still solves for δ. That both the Lanave *et al.* (1984), and the Rodriguez *et al.* (1990) distances are based on the same model has been noted by Zharkikh (1994). Interestingly when assisting David Swofford to implement this distance into PAUP*, we noticed these two formulae give exactly the same distance (to 12 decimal places), suggesting their algebraic identity.

Another interesting feature of the solution of the time reversible model is that it not only recovers the total distance, but it also infers the underlying matrix of relative rates of change (**R**t). Given our estimate of π, we can very simply estimate the total numbers of changes of different types. For example number of A $\rightarrow$ C changes = $\pi_a r_{ac} t$, A $\rightarrow$ G changes = $\pi_a r_{ag} t$, i $\rightarrow$ j changes = $\pi_i r_{ij} t$ (as Lanave *et al.* 1984 do). While there appear to be twelve independent rates there are in fact only 9 underlying rates, so that some entries in **R** may be considered products of an underlying rate multiplied by the frequency of a state, i.e. Π (see Tavaré 1986)(note these same results can easily be extended to any number of states, for example the twenty amino acids). Estimates of these relative rates should be of general interest to biologists trying to understand the mutation / substitution process (and potentially much more informative than the 3 rates of the Kimura 3ST model, for example). Of course we may have good reason to doubt that the relative rates of substitution form a time reversible process, in which case we may wish to model a general 12 parameter model, and recover the 12 rate categories this way. For four states this can only be done using a predictive model (e.g. maximum likelihood, or minimum X^2), analysing 3 or more sequences simultaneously (e.g. see Blaisdell 1985). Predictive models, by using all the information simultaneously, are also expected to be a more efficient way of

obtaining estimates of the overall rates than a multitude of non-independent distance comparisons.

If we are assuming a time reversible model, then we expect all divergence matrices to be symmetric. A proof of this expectation is given in appendix 3.2 (developed with Dr Mike Steel). Consequently we can reduce the error of any distance measure made under a time reversible model if we can make a joint estimate of $f_{ij} = f_{ji}$. As Barry and Hartigan (1987b) mention, this amounts to forming a symetricised divergence matrix, and this matrix **F*** is also the ML estimator of **F** (since $\hat{\mathbf{F}}$ is expected to have a multinomial distribution, then the ML estimator of $f_{ij} + f_{ji} = \hat{f}_{ij} + \hat{f}_{ji}$, and since $f_{ij} = f_{ji}$, it follows that the ML estimator of f_{ij} or f_{ji} is $1/2\,(\hat{f}_{ij} + \hat{f}_{ji})$). Since **F*** is the ML estimator of $\hat{\mathbf{F}}$, and $\hat{\Pi}$ is the ML estimator of Π (and since **F*** is the sufficient statistic under this model) then it follows that equation 3.2.1-1 is also the ML estimator of the distance under the general time reversible model. The expectation that **F** will be symmetric, also allows a test of whether the data were generated by a time reversible substitution scheme (although when comparing more than one pair of distances their covariances need to be taken into account in a similar manner to Bulmer 1991a, for example).

3.2.2 A distribution of rates across sites with the general time reversible distance

We can also extend distances estimated under stationary time reversible models (with up to 9 parameters in their transition matrices) using the same general approach used in chapter 2 with the Hadamard conjugation. Our new distance estimation formula estimating the expected number of substitutions per site is,

$$\delta_{ij} = -\text{trace}[\Pi M^{-1}(\Pi^{-1}\mathbf{F})] \qquad (3.2.2\text{-}1)$$

where M^{-1} is the inverse of the moment generating function of the distribution of rates across sites (table 2.2 gives some examples). The application of M^{-1} to $\Pi^{-1}\mathbf{F}$, (here taken as matrix **Z**) is defined as,

$$M^{-1}(Z) = \Omega M^{-1}(\Psi)\Omega^{-1} \qquad (3.2.2\text{-}2)$$

where Ω is a matrix containing as columns the eigenvectors of **Z**, Ω^{-1} is its inverse, and function M^{-1} is applied componentwise to the diagonal matrix Ψ which contains the eigenvalues of **Z**. As with the time reversible i.r. model we symmetrise **F**, as doing this we are always able to diagonalise the product of $\Pi^{-1}\mathbf{F}$ (the product of two symmetric matrices always has real eigenvalues). This method also provides a quick way of calculating **P** across any edge or path in a tree if rates at sites vary. Let, w_{mn} be equal to the total expected number of substitutions on an edge or along a path, while **R** is a normalised Markov instantaneous rate matrix such that its positive entries sum to 1, then

$$\mathbf{P} = M(\mathbf{R}w_{mn}), \qquad (3.2.2\text{-}3).$$

This last result allows us to quickly calculate the divergence matrix under any continuous times Markov process where sites evolve independently, and the transition matrix can be written in the form $\mathbf{R}\lambda_i$ (where is the relative rate of substitution at the *i*-th site). Again if **R** defines a time reversible process it can be diagonalised and has real eigenvalues (Keilson 1979, section 3.2). A proof of the last three equations is similar to that for Hadamard conjugations with a distribution of rates across sites, and relies upon the fact that $\mathbf{F} = \Pi\Sigma\exp(\mathbf{R}t\lambda_i)$, and will appears in an manuscript with Dr Mike Steel (Waddell and Steel unpublished). That **F** is a linear sum of time reversible processes, guarantees complete monotonicity of time reversible distances with a distribution of rates across sites. Further we can derive the variance of this estimate using a delta method approximation (see Barry and Hartigan 1987b), and this gives:

$$Var[\hat{\delta}] \approx \frac{1}{c}\left[\sum_{k=1}^{4}\pi_k(R_{kk}-\sum_{i=1}^{4}\pi_i R_{ii})^2+\sum_{k=1}^{4}\pi_k\left\{\sum_{l=1}^{4}P_{kl}\left(G_{kl}-\sum_j P_{kj}G_{kj}\right)^2\right\}\right]+O(c^{-2}) \quad (3.2.2\text{-}4)$$

where $(\mathbf{G}_{kl})$ are elements of the matrix, $\mathbf{G}=-\sum_{r=1}^{\infty}a_r\sum_{s=0}^{r-1}B^s(B^t)^{r-1-s}$, $\mathbf{B} = \mathbf{I} - \mathbf{P}$ (where t indicates transpose). The term a_r is equal to $M^{-1}(1\text{-}x) = \sum_i a_i x^i$. For example, in the case of the Γ distribution this gives,

$a_n = [(k+1)(2k+1)...((n-1)k+1)]/(n!k^n)$, e.g. with shape parameter $k = 0.8$, $a_1 = 1/0.8 = 1.25$; $a_2 = (0.8+1)/(2!(0.8^2)) = 1.41$; $a_3 = [(0.8+1)(1.6+1)]/(3!(0.8^3))$, etc.

However the proof in Barry and Hartigan (1987b) appears to contain some errors which in collaboration with Mike Steel, we are looking to correct if necessary, then check the accuracy of this approximation with simulations. By a manipulation of this formula it is also possible to estimate the variance-covariance matrix of the entries (rates) in **R** (not shown).

Being able to predict **P** for such a general model, allows us to look into the consistency (and for large samples the robustness and convergence) of tree building methods based on specific assumptions about **R**, Π or the distribution of rates across sites (we can even make **F** a sum of different **F** matrices, and so model different sets of sites of sites having a different underlying **R** matrix).

A second use of the method is to estimate pairwise distances under any i.i.d. model. We can predict **F** for any homogeneous and stationary i.i.d. model (requiring **R** and t, or t_1 and t_2 if the process is not time reversible), then it is straight forward to fit **F** to $\hat{\mathbf{F}}$, and the estimated distance is that which maximises the fit (Felsenstein 1993, uses this approach in PHYLIP to make a maximum likelihood estimate of distance under the stationary K 2ST model). It is worth noting that Blaisdell (1985) suggested a similar procedure to infer distance plus parameters, something which can perhaps best be done with ML or X^2 fitting of **s**(T)(inferred sequence patterns across all taxa) to data. The real advantage of this type of distance estimation is that if we assume a stationary model, then we can fix **R** so that distances estimates are more efficient

(i.e. remove some stochastic variation). This in turn is expected to help in reduce sampling variance in both distance estimation and tree selection from these distance estimates (e.g. Kuhner and Felsenstein 1994).

Let us assume that we have evidence that a stationary time reversible model (where **R** is of the form **S**Π as previously defined in section 3.2.1) is a reasonable fit, and we have estimates of the nine independent entries in **S** and Π. Further assume that analyses like those illustrated latter in the chapter suggest that the distribution of rates across sites is well approximated by a gamma distribution with shape parameter k = 0.8. Further we expect that the rate of A ↔ C is equal to that of G ↔ T in the symmetrical matrix **S**, and the base composition is consistent with the frequency of G and C being equal. Thus we average these specific entries and are ready to infer δ, since we have **S**Π = **R** (with 9-2 = d.f.), i.e.

$$\underset{\mathbf{S}}{\begin{bmatrix} - & \alpha & \beta & \varepsilon \\ \alpha & - & \phi & \eta \\ \beta & \phi & - & \alpha \\ \varepsilon & \eta & \alpha & - \end{bmatrix}} \times \underset{\Pi}{\begin{bmatrix} \pi_a & 0 & 0 & 0 \\ 0 & \pi_c & 0 & 0 \\ 0 & 0 & \pi_c & 0 \\ 0 & 0 & 0 & \pi_t \end{bmatrix}} = \underset{\mathbf{R}}{\begin{bmatrix} - & \alpha\pi_c & \beta\pi_c & \varepsilon\pi_t \\ \alpha\pi_a & - & \phi\pi_c & \eta\pi_t \\ \beta\pi_a & \phi\pi_c & - & \alpha\pi_t \\ \varepsilon\pi_a & \eta\pi_c & \alpha\pi_c & - \end{bmatrix}}$$

(we require estimates of all parameters in **R**, estimating them if we need to from each pairwise **F** matrix, but preferably with a method such as ML based on the sequences of many taxa). It then follows that **F** = Π(M(**R**t)), where **R** is given, and M is applied to the whole matrix (via diagonalisation as described earlier), and in this case M is equal to $((k - x) / k)^{-k}$, or $((0.9 - x) / 0.9)^{-0.9}$ (see table 2.2). It just remains to alter t until a best fit to is obtained between $\hat{\mathbf{F}}$ and **F** (or **F*** and **F**) and the final t is our distance estimate (note we must normalize **R** so that non-diagonal elements of each row sum to 1 if we wish t to be the expected number of substitutions per site and not a multiple of this quantity).

This approach also allows the estimation of distances under a model we anticipate will be a especially useful for analysing "neutral" nuclear DNA (e.g. psuedogenes). An ongoing conjecture in molecular evolution is whether neutral substitution rates on each strand of nuclear DNA are equal. Various studies have looked at this question using statistical tests and concluded that there is no good evidence to refute this hypothesis (e.g. Bulmer 1991b), making it seem likely to be an accurate approximation to reality. Taking these studies one step further, we can infer the rate matrix under such a model (a new step as far as I am aware). Due to canonical base pairing of DNA, the constraint of equal substitution rates on each strand implies a set of six constraints (leaving a model with 6 degrees of freedom). Because A pairs with T and C with G, this implies, for example that the rate of the change A → T must equal the rate change of T → A, or that the rate A → C must equal the rate of T → G, etc. This specifies the rate matrix,

$$\begin{array}{c} \\ A \\ C \\ G \\ T \end{array} \begin{array}{c} \begin{array}{cccc} A & C & G & T \end{array} \\ \begin{bmatrix} - & a & b & c \\ a & - & c & b \\ d & e & - & f \\ e & d & f & - \end{bmatrix} \end{array}$$

This rate matrix is non-time reversible. So we must infer $\mathbf{F}_{mn} = (\mathbf{P}_{rm})^t \Pi \mathbf{P}_{rn} = (M(\mathbf{R}t_1))\Pi(M(\mathbf{R}t_2))$, and our distance is $(t_1 + t_2)$, where t_1 is the distance (assuming **R** is normalised) from the root (r) of this path to sequence m, and t_2 is the distance from the root to n. This rate matrix does have a closed form expression for its eigenvalues (M. Waddell pers comm.), so it may be possible to obtain a closed form expression for the true distance. It seems unlikely that any closed form expression would be either an ML estimator or would easily allow the underlying rate matrix to be fixed for all sequence comparisons. Thus this computationally more expensive "fitting" approach may be preferable. With neutral DNA M would usually be taken to be the moment generating function of the delta function (i.e. identical rates at all sites). However during ML fitting of data to model to make the divergence time estimates in Waddell and Penny (1995), it was found that a distribution of rates across sites could improve fit of psuedogene data to model (although not significant at the 5% level; more details are given in chapter 5). Thus we should keep an open mind on this issue and use a non-i.r. transformation if there is good evidence for site to site rate variation.

3.3 DISTANCE ESTIMATION UNDER NON-STATIONARY MODELS

Here we look at estimators which are hopefully robust to the effect of sequences having quite different nucleotide base compositions, a problem recently recognised as potentially serious trouble for tree estimation (see for example Weisburg *et al.* 1989, Lanave *et al.* 1989, Lockhart 1990, Loomis and Smith 1990, Lockhart *et al.* 1992, Hasegawa *et al.* 1992). Here we describe some new properties of LogDeterminant (also called "Paralinear distances) which have recently been suggested as a major breakthrough in helping to combat such biases (see especially Lockhart *et al.* 1994). We also look at another class of approximate corrections which hopefully make most distance transformations more robust to non-stationary base compositions.

3.3.1 LogDet distance methods including new results on their interpretation

While we cannot recover a distance δ as the number of substitutions per site for the general non-stationary model, there is a way of obtaining a tree additive distance under any i.i.d. and i.r. Markov model such that,

$$\delta_{mn} = -\ln[\det(\mathbf{F}_{mn})] \qquad (3.3.1\text{-}1)$$

where det is the matrix determinant (Steel 1994a). Subsequently Lockhart *et al.* (1994) suggested the formulae,

$$\delta_{mn} = -1/4\ (\ln[\det(\mathbf{F}_{mn})] - 1/2\ \ln[\det(\Pi_m)] - 1/2\ \ln[\det(\Pi_n)]), \qquad (3.3.1\text{-}2)$$

(the same formula was independently derived by Lake 1994, but leaving out the factor 1/4 which as we will see is useful to include). For explicit mathematical proofs of these distances under this general models, see Steel (1994a), and Steel *et al.* (1993a).

Let's look at what these equations are doing. Recalling that $\mathbf{F}_{mn} = \mathbf{P}^t_{rm}\Pi_r\mathbf{P}_{rn}$, so $\det(\mathbf{F}_{mn}) = \det(\mathbf{P}^t_{rm}\Pi_r\mathbf{P}_{rn}) = \det(\mathbf{P}^t_{rm})\det(\Pi_r)\det(\mathbf{P}_{rn})$ (a well known result in matrix algebra, which also states that the order of multiplication of determinants is commutative and associative). So $-\ln(\det(\mathbf{F}_{mn})) = -\ln(\det(\Pi_r)) + -\ln(\det(\mathbf{P}^t_{rm}\mathbf{P}_{rn}))$. Further the last term can be written as $-\ln(\det(\mathbf{P}_{mn}))$, and so $-\ln(\det(\mathbf{P}_{mn})) = -\ln(\det(\mathbf{P}_{m1})) + -\ln(\det(\mathbf{P}_{12})) + ... + -\ln(\det(\mathbf{P}_{kn}))$ (where k is the label of the last internal node on the path from m to n) (see Cavender and Felsenstein 1987, Barry and Hartigan 1987b, Steel 1994, and Lake 1994 for a mathematical description of this property applied to tree based sequence evolutionary models). The taking of logarithms has converted a multiplicative property of a distance to an tree additive property, just the kind of distance we need for most distance based tree building algorithms to be consistent (Felsenstein 1988). Note that under any stochastic process the determinant of **P** will always be a number greater than zero, unless it indicates completely random change (in our case indicating an infinitely long edge in the tree) where it will take value zero. Consequently barring sampling error, its log will always be defined under the model. Likewise for any Markov process, the determinant of Π will always be between zero and one (being zero only if one of the states has frequency zero, which should only happen due to sampling error, as any process with all **P** matrices having positive determinants will also generate Π matrices having positive determinants).

It is important to understand what the LogDet formula is measuring. If $\mathbf{P}_{mn}$ is divided up into k equally spaced intervals, then as k tends to infinity each interval becomes Δt, and the transition matrix in one of these very fine intervals, $\mathbf{P}_{\Delta t}$, becomes equal to $(\mathbf{I} + \mathbf{R}\Delta t)$(that is the identity matrix, **I**, plus the instantaneous rate matrix, **R**, for the instant Δt). In turn, $\ln[\det(\mathbf{I} + \mathbf{R}\Delta t)]$ becomes equal to ln[det(1 + trace(**R**)] (since det of **I** = 1, and det of **R** as k tends to infinity is trace of **R**), which in turn becomes -trace(**R**) (as the logarithm of a number becoming barely less than 1, tends to ln(1) (which is 0) minus that number). So, the quantity -trace(**R**) can be interpreted as four times the sum of all the nucleotide changes, $i \rightarrow j$, divided by the frequency of nucleotide i, in that instance of time. Compare this with the formula introduced earlier which said that the total number of substitutions per site was -trace(Π**R**), and you can see that the two would be the same if we multiplied by Π^{-1} (i.e. $\mathbf{R} = \Pi^{-1}\Pi\mathbf{R}$). Multiplying by the diagonal matrix Π^{-1} is the same as dividing all elements in Π**R** by $1/\pi_i$, which is exactly the additional weighting factor we have coming in when we consider what trace of **R** is counting. So while more common distance measures such as the Jukes-Cantor, and even the general time reversible distances, aim to recover the unweighted sum of all substitutions, $\sum_{i=1}^{4} \Pr(i \rightarrow j), i \neq j$ (where i $\rightarrow$ j are changes such as A $\rightarrow$ C), in contrast minus 1/4 of the logarithm of the determinant of **P** is summing up

$\sum_{i=1}^{4} \frac{\Pr(i \rightarrow j)}{frequency(i)}, i \neq j$, where the frequency of i is estimated in the instant of time that any change is occurring. If the process is non-stationary (with respect to the base frequencies), then the frequency of state i will change as we move along edges of the tree. This interpretation of what -ln(det($\mathbf{P}_{mn}$)) is counting is due to Barry and Hartigan (1987b), and the example above loosely follows their proof.

There are is a set of special cases where -1/4ln[det($\mathbf{P}_{mn}$)] will recover the distances from m to n as the number of substitutions per site. Any model where the frequency of all four bases stays at 0.25, will meet this condition (it is here that the 0.25^{-1} weighting of each substitution is canceling with the 1 / 4 factor at the front of the equation to give exactly the number of substitutions per site, Barry and Hartigan 1987b, Lake 1994, and Lockhart *et al.* 1994). Since -1/4ln[det($\mathbf{F}_{mn}$)] is equal to -1/4ln[det($\mathbf{P}_{mn}$)] + -1/4ln[det(Π)], then this also explains why Lockhart *et al.* (1994) suggested a version of the LogDet which is -1/4ln[det($\mathbf{F}_{mn}$)] -1/4rln(r) (where r is the number of states, here 4) since the last term rln(r) is exactly equal to -1/rln[det(Π)] for any equi-frequency model with r states. Equi-frequency models include all models with symmetric rate matrices, and a special subset of asymmetric rate matrices.

Under any stationary Markov model, then the most general LogDet distance (equation 3.3.1-2) will be linear with the number of substitutions per site (and with a molecular clock linear with time). This makes it potentially very useful for constructing weighted trees, when we aim to calibrate one or more nodes in the tree with a specific age (e.g. from fossil evidence) and then infer the age of other nodes by ratios of edge lengths. Using the LogDet transformation rather than any of the more specific distance transformations frees us from any specific assumption about the form of **R**. (Note that for this purpose all other pairwise distance measures will give systematic underestimates if **R** contains 12 parameters, or 10 or more parameters if a molecular clock holds, since **F** must then be symmetric). Since **F** is expected to be symmetric under a molecular clock (proof in appendix 3.2), then if we are imposing a clock to hold over the whole tree, we can reduce the variance of the LogDet distance by symmetrisation (i.e. **F*** as already discussed in section 3.2.1 on general time reversible distances). The same argument holds if we wish to constrain the LogDet equation to a time reversible model, where we should also apply the LogDet to **F*** (which also must have a symmetric **F** matrix, see appendix 3.2).

Results from the previous paragraph also suggest a new form of LogDet if we expect that the base frequencies are stationary, but not necessarily equal (nor need the model be time reversible). In this case there should be a Π matrix common to all sequences, and it is more accurately estimated as the average of all sites in the sequences (not just a pair at a time). We denote this estimate as Π_{av} (average over all sequences), so out additive distance measure becomes,

$$\delta_{mn} = -1/4\ (\ln[\det(\mathbf{F}_{mn})] - \ln[\det(\Pi_{av})]). \qquad (3.3.1\text{-}3)$$

Equation 3.3.1-3 is additive on a tree under the same assumptions as other LogDet measures since $\ln[\det(\mathbf{F}_{mn})]$ is additive (Steel 1994a) and $-\ln[\det(\Pi_{av})]$ is a constant added to all distances thus not altering this additivity (Barry and Hartigan 1987b). (It is possible to constrain **F** to **F*** to make this estimate under a clock or a stationary time reversible model). Due to sampling errors and violations of the assumption of stationarity 3.3.1-3 could return a negative distance between closely related species (especially if their base composition was closer to equi-frequency than the average of the sequences used to estimate Π_{av}). If using these distances to estimate a tree with an algorithm that required all distances greater than or equal to zero, it may be necessary to add a small constant to all pairwise distances so that the smallest distance became zero. Its not clear if using Π_{av} will substantially reduce the variance of LogDet distances in this type of situation. We presently understand little of the variances and covariances of different LogDet distances, and especially how these alter the reliability of tree selection when coupled with various tree selection criteria and algorithms.

It is interesting to note the connection between the LogDet of the **F*** matrix and the time reversible distance of equation 3.2.1-1 applied to **F***. Specifically,

$$\delta_{ij} = -\text{trace}(\ln[\Pi^{-1}\mathbf{F}^*]) = -\text{trace}(\mathbf{R}t) = -(\ln[\det(\mathbf{P}^*)]) = -(\ln[\det(\mathbf{F}^*)] + \ln[\det(\Pi^{-1})]), \quad (3.3.1\text{-}4)$$

where by definition **R** is a the rate matrix which gives rise to a time reversible model (**F***) and **P*** is **P** estimated from **F***. This relationship holds by the Jacobi identity which states that the determinant of a matrix is equal to the product of its eigenvalues (and because any symmetric matrix has all real eigenvalues, and gives rise to a unique **R** matrix). With real data both **F*** and Π are estimated as described earlier. Thus we suspect that the sampling variance of these additive distance estimates is also equal. Under the Kimura 3ST model (and submodels) a further restriction is also possible, so that all entries in **F*** which are expected to be equivalent are averaged (e.g. all types of transitions equal, all type 1 transversions equal, all type 2 transversions equal, and all diagonal elements equal). If we call this matrix $\mathbf{F}^{\#}$, then it follows that,

$$\delta_{ij} = -\text{trace}[\ln(\Pi^{-1}\mathbf{F}^{\#})] = -\ln(\det(\mathbf{P}^{\#})) = -\text{trace}(\mathbf{R}t) = -(\ln(\det(\mathbf{F}^{\#})) - 4\ln(4)) \quad (3.3.1\text{-}5)$$

More specifically under the Kimura 3ST model, then $-\text{trace}(\mathbf{R}t) = 4 \times$ (no. substitutions per site) $= -[\ln(1\text{-}2P\text{-}2Q) + \ln(1\text{-}2P\text{-}2R) + \ln(1\text{-}2R\text{-}2Q)]$. This last equation, which is Kimura's (1981) equation bar the factor 1/4, is thus equal to the constrained LogDet equation, i.e. $\ln(\det(\mathbf{P}^{\#}))$. This relationship holds because the terms 1, (1-2P-2Q), (1-2P-2R), and (1-2Q-2R) are the eigenvalues of $\mathbf{P}^{\#}$. Thus this constrained form of the LogDet is another way of writing the 3ST equation. Consequently, the sampling variance of the constrained LogDet under the Kimura 3ST model and its submodels, must be equal to that of the standard ML estimators under these models. (Note that the standard Jukes-Cantor, Kimura 2ST and Kimura 3ST distance formulae are ML distance estimators under the Poisson, generalised Kimura 2ST and generalised Kimura 3ST models respectively, e.g. see Zharkikh 1994). This helps to explain why even the unconstrained LogDet transformation has a very similar variance to the ML estimators of δ_{ij}

under the Kimura 3ST model (and submodels, unpublished simulations with David Swofford)(also see chapter 4 where it is shown that the delta method variance approximations of these two methods are equal).

Unfortunately, even in these constrained cases of the LogDet, we cannot allow for a distribution of rates across sites without first separating sites into rate classes. Consider for example that $\ln[\det(P_{mn})] = \ln[\text{product of eigenvalues of } P_{mn}] = \ln[e_1] + \ln[e_2] + \ln[e_3] + \ln[e_4]$ which is then an additive measure due to the instantaneous interpretation (detailed below) that this is a measure of the sum of the trace of the **R** matrix at each instant of evolution (if we assume a continuous process). It is not possible to replace *ln* with a moment generating function and get the analogue of the time reversible distance with a distribution of rates across sites (3.2.2-1), since $M^{-1}[\det(P_{mn})] \neq M^{-1}[e_1] + M^{-1}[e_2] + M^{-1}[e_3] + M^{-1}[e_4]$. Others ways of improving additivity with unequal rates across sites must be found.

An interesting new result we add here is that under any stationary i.i.d. and i.r. Markov model, the expected value of equation 3.3.1-2 is always greater than or equal to the total number of substitutions. This amounts to showing that with Π in equilibrium with **R**, then $-1/h \times \text{trace}(\mathbf{R}) \geq -\text{trace}(\Pi\mathbf{R})$ (where h is the number of character states). We have evaluated this condition with 50 twelve parameter **R** matrices of very different forms plus random numbers and it was found to hold in all cases (in addition the amount that -trace(**R**) exceeded the number of substitutions per site was highly correlated with how far Π was from being equifrequency). (I am presently working on a proof of this conjecture with Mike Steel, so far we have proved this conjecture for 2 character states. The general proof is elusive because we have yet to find a closed form expression for the equilibrium base composition under the general i.i.d. model with more than two states).

Lastly now, how do we interpret the LogDet distance (equation 3.3.1-2) under the most general i.r. / i.i.d. model of evolution? A straight forward way of doing this is to realise that,

$$-1/8\,(\ln[\det(\mathbf{P}_{mn})] + \ln[\det(\mathbf{P}_{mn})]) = -1/4\,(\ln[\det(\mathbf{F}_{mn})] - 1/2\,\ln[\det(\Pi_m)] - 1/2\,\ln[\det(\Pi_n)]), \quad (3.3.1\text{-}5)$$

(a proof of this is presented in appendix 3.3). As we will demonstrate in section 3.4.2 the asynchronous distances of Barry and Hartigan (1987b) are not tree additive under a non-stationary model, however the average of δ_{ij} and δ_{ji} is additive and identical to the favoured form of LogDet used by Lockhart *et al.* (1994), and Lake (1994) (appendix 3.3). So combining this result with our previous result we obtain,

$$-1/4\,(\ln[\det(\mathbf{F}_{mn})] - 1/2\,\ln[\det(\Pi_m)] - 1/2\,\ln[\det(\Pi_n)]) =$$

$$\frac{1}{8c}\left(\sum_{i=1}^{4}\frac{\Pr(i \rightarrow j)}{frequency(i)} + \sum_{j=1}^{4}\frac{\Pr(j \rightarrow i)}{frequency(j)}\right), \quad i \neq j \qquad (3.3.1\text{-}6)$$

(where frequency refers to normalised frequency or proportion). If we multiply our the LogDet distance of equation 3.3.1-2 by the sequence length we have the following interpretation: This distance is the average of, the number of substitutions $i \rightarrow j$ weighted by the initial state (i), plus the number of substitutions $i \rightarrow j$ weighted by the state it changed into (j). For example, assume a change A $\rightarrow$ C occurred at point x between tips m and n of the tree, and the frequency of A was (0.1) at this instance, while the frequency of C was (0.4). This change contributes 1/8(1/0.1 + 1/0.4) = 1/8(10 + 2.5) = 1.5625 to the distance between points m and n. Notice that this is not the same as 1/4(1 / [(f(a) + f(c)) / 2], which gives 1/4(1 / [(0.1 + 0.4) / 2]) = 1, so there is no general way to simplify this distance interpretation. Consequently, the distance in (3.3.1-6) is tree additive in expectation (so can be expected to give consistent tree estimation when combined with a reliable tree building algorithm under this i.r. model). As we will see below this distance can also be a useful approximation to the more traditional distance of the unweighted number of substitutions per sites under non-stationary conditions (then often being a better estimator of this quantity than the standard distances, e.g. JC, K 2ST, etc.).

Lastly notice that interpretation 3.3.1-6 of the LogDet equation shows that there is no arrow of time in this distance, it is gives identical values in either direction. Also contrary to the expectation of Barry and Hartigan (1987b) there is no special information in their asymmetric distance. The difference in two asynchronous distances is just $\ln[\det(\Pi_m)] - \ln[\det(\Pi_n)]$, a rather insensitive measure of inequality of base compositions (e.g. it can be zero if base compositions are quite different, but the diagonal of the two Π matrices are permutations of each other). Its most general feature is that it is always positive if the base composition of m is equifrequency and that of n is not. Consequently $\ln[\det(\Pi_m)] - \ln[\det(\Pi_n)]$ offers no special information and would only be useful in locating the root when assuming the ancestral base composition was [0.25: 0.25: 0.25: 0.25].

Before we move on it is worth noting that both the general distances (the extended time-reversible distance with unequal rates across sites and the LogDet distance) can be applied to any number of states (e.g. Lockhart *et al.* 1994, Lake 1994). In coding regions it is often the amino acid and not so much the individual bases in the codon, that acts as a single unit. Consequently a rate matrix for amino acid changes may better approximate the course of evolution than a single rate matrix applied to all sites in the DNA sequence (or even first, second, and third DNA positions treated separately) (e.g. this is the reason for PROTML, Adachi and Hasegawa 1992, see also Reeves 1992). However a further refinement is that even amino acid substitutions, while being selected for by the properties of the end product (the protein), are driven at different rates depending on what the underlying codon is, and how readily a particular mutation to another codon can occur (a property of the structure of the genetic code, and the rates of nucleotide base mutation). Consequently it is expected that a codon to codon transition matrix (64 possible states, with the 3 stop codons typically excluded), may even better describe the evolution of a stretch of protein coding DNA (Muse and Gaut 1994, and Goldman and Yang 1994 have recently studied this question with an ML model). Thus it may be fruitful to apply either time reversible distances, or LogDet, to a 61 $\times$ 61 **F** matrix of the divergence of codons

between two sequences, or to a 20×20 matrix of the divergence of amino acids (or alternatively LogDet **F*** if we assume a time reversible model, which will also help reduce sampling variance). Whether either of these approaches can become generally used will depend largely on the kinds of sampling error (variance and bias) that can arise when using typical sequence lengths of 200 to 2000 base pairs. With so many possible pairs of states, and relatively short sequences, it is expected many of the rarer events in the divergence matrix will not be observed, and there will also sometimes be completely missing states, e.g. a short protein with no cystine residues. This will require either nominal values in $\hat{\Pi}$ or $\hat{\mathbf{F}}$, some pooling of states such as into amino acid classes, or else exclusion of a missing state in order to give non-zero determinants and thus defined distances. It would be most desirable to assess the reliability of these distances (based upon amino acid or codon changes) using simulations before they were used regularly to analyse real data.

One further extension may be of limited use with the LogDet method. It is desirable to be able to model insertion-deletion events as well as substitution events, when analysing DNA sequences. If an insertion-deletion is more than one position in length (or one codon if using this as the state) then it clearly violates the i.i.d. assumption, as applied to the other states being analysed. There is also the issue of deletions being possible absorbing states, when there can be change into a state but not out of it. However there may be scope for including the shortest insertion deletion events (say 2 sites or less in length) into the divergence matrix, in the hope that the additional information is reliable enough to overcome the model violation which is occurring. Non-coding regions can, for example, show a high rate of very short insertions and deletions. Again simulation results from a realistic model should be most illuminating regarding the utility of such distance estimates.

It is worth noting that the LogDet transform is to phylogenetic analyses based on distances, what the general i.i.d. and i.r. model of Barry and Hartigan (1987a) is to maximum likelihood tree inference from sequences. This has several important consequences not previously noted. Firstly, evidence that LogDet distances are not fitting a tree building model reliably, also implies that the general i.i.d. ML model will not explain the same data adequately. This could be taken as proof of the need to consider more sophisticated models, perhaps incorporating correlations between sites or elements of high order molecular structure and function. Secondly, just as there is no obvious way to extend the LogDet to exactly model unequal rates across sites (without separating them into rate classes), then the same restrictions apply to the general i.i.d. ML model. The root cause of this is that the general i.i.d. ML model allows for any **P** matrix with positive determinant on an each edge of the tree. However this **P** matrix need not correspond to any homogenous continuous time process, so there is no direct way to infer what the corresponding **P** matrix would be if there were unequal rates across sites (e.g. for a single edge the relationship of 3.2.2-3 no longer holds).

Previously the LogDet has not been described as an ML distance estimator, however it probably is. For example the LogDet applied to $\mathbf{F}^{\#}$ (the **F** matrix formed by averaging entries which have equivalent likelihood under the equilibrium root distribution Kimura 3ST model or

submodels) reduces to the more familiar ML estimators of Jukes and Cantor (1969), Kimura (1980), and Kimura (1981) if base composition is assumed to be equifrequency. Also the LogDet applied to **F*** (the symmetrised matrix) gives equivalent estimates to the ML general time reversible distance, except that it is an ML estimate of the weighted distance (3.3.1-6). Lastly $\hat{\mathbf{F}}$ is the sufficient statistic under the general i.r. / i.i.d. model, so it seems likely that it too is an ML distance estimator.

To apply the general i.i.d. model reliably when sites do not have identical rates, either an invariant sites model must be used, or else sites must be classified into separate rate classes. A third point of note is that it will be interesting to compare distance based methods (LogDet) and ML methods of tree selection under this very general model with short sequences. The ML model is estimating 12 (2t-3) parameters per tree, whereas the LogDet is apparently only estimating one thing, namely the determinant of the **F** matrix. ML methods are expected to do better than distance based methods with long sequences but it would be very interesting if ML did better than distance based methods even with short sequences (say 100 or more sites, with four or five taxa) under the general i.r. - i.i.d. model. This in turn would beg the question of how a ML method which is apparently estimating so many parameters, can do better with very limited data than an apparently more simple distance method estimating a only a few key quantities.

This section is concluded with a look at the determinant of a 4 × 4 matrix. We do this because it is interesting to consider how sampling error could bias this estimate of a determinant, and to notify the reader of a computational error that was detected. The determinant of a 4 × 4 matrix can be written as:

$$\det(F) = dt1 - dt2 + dt3 - dt4,$$

where

$$dt1 = f_{11}\{ f_{22}(f_{33} \times f_{44} - f_{34} \times f_{43}) - f_{23}(f_{32} \times f_{44} - f_{34} \times f_{42}) + f_{24}(f_{32} \times f_{43} - f_{33} \times f_{42})\},$$

$$dt2 = f_{12}\{ f_{21}(f_{33} \times f_{44} - f_{34} \times f_{43}) - f_{23}(f_{31} \times f_{44} - f_{34} \times f_{41}) + f_{24}(f_{31} \times f_{43} - f_{33} \times f_{41})\},$$

$$dt3 = f_{13}\{ f_{21}(f_{32} \times f_{44} - f_{34} \times f_{42}) - f_{22}(f_{31} \times f_{44} - f_{34} \times f_{41}) + f_{24}(f_{31} \times f_{42} - f_{32} \times f_{41})\},$$

$$dt4 = f_{14}\{ f_{21}(f_{32} \times f_{43} - f_{33} \times f_{42}) - f_{22}(f_{31} \times f_{43} - f_{33} \times f_{41}) + f_{23}(f_{31} \times f_{42} - f_{32} \times f_{41})\}.$$

Evaluations to date show that with short sequences there is a slight downward bias to the expected value of the determinant estimated in this way. Overall bias is then increased by the non-linear log transform, so that with short sequences the LogDet tends to overestimate the asymptotic distance (unpublished simulations with D.L. Swofford and P.O. Lewis, and unpublished analyses by D. Penny).

Later in section 3.7 we do some extensive analyses of 28 16S-like rRNA sequences using the LogDet equation 3.3.1-2, as implement by David Penny in the programs "Prepare" and "Trees" (Penny *et al.* 1993, available by anonymous FTP from FARSIDE@massey.ac.nz). A programming error has recently been detected in both these programs, and although now corrected (as of July 1995) there has not been time to rerun and redraw all the analyses. The error occurred in estimating the determinant of **F** as dt1 - dt2 + dt3 + dt4 when the last term

should have been subtracted. For many distance estimations this error makes a difference only in the third (or fewer) significant place. The reason is that this last dt term tends to be the smallest of the four (as the **F** matrix is programmed the first term of dt4, f_{14}, is the relatively rare transversion A $\rightarrow$ T). At higher rates of change this error results in the determinant being overestimated, and in fact seems to act to counter the bias of the standard distance in some situations. In our case we have been able to check the main results (figures 3.11 and 3.12) using early test versions of the program PAUP* which was giving correct distances. Even with the large amounts of change in these sequences, and with the removal of invariant sites, this programming error did not change any directly evaluated tree, and made only minor differences to bootstrap support estimates (generally less than 5%, and these differences are noted in the relevant text). Consequently we expect that the published analyses made using the programs "Prepare" and "Trees" to estimate LogDet distances will not be misleading. We have also mentioned this point partially because this programming error may actually point the way towards unbiased estimators, and also hints at a fair degree of robustness in this general formula. All other distance estimates were made with the program PHYLIP (Felsenstein 1993), or PAUP* (Swofford in press).

3.3.2 Approximate methods to give robustness with varying base composition

There have been intermittent claims in the literature of non-LogDet methods which are robust to non-stationary base compositions in estimating additive distances. Here we investigate these modified "stationary distances" and show they can be applied to a wide range of distance transformations. Whether they actually help (relative to standard distances) is an open question not yet considered theoretically. One possibility to increase the robustness of distance transformations is to modify them so that they return an "infinite distance" (e.g. taking the logarithm of zero) when we observe data which we expect to be random when taking into account the observed base frequencies in each sequence. This observation is used to modify Kimura 2ST and Kimura 3ST distances, and to consider such distances properties. Examples of this approach have been given by Olsen (reported in Weisburg *et al.* 1989) and Bulmer (1991a) in the instance of the Jukes-Cantor, and "equal input" models (e.g. Felsenstein 1981a, Tajima and Nei 1984, Tavaré 1986). For example the Jukes-Cantor distance $\delta = -3/4 \ln(1-4/3d_{obs})$ assumes that at very large distances the observed distance for long sequences will approach 3/4, and hence the logarithm of the term in the brackets will approach infinity. If base compositions are not equal in the two sequences, we could expect random sequences to show either more or less than 1/4 matches when randomised.

A more general distance correction that the Jukes-Cantor, often called the "equal-input" model (Tajima and Nei 1984), has a very similar distance correction formula, $\delta = -b \ln[1-d_{obs}/b]$, where $b = (1 - f_a^2-f_c^2-f_g^2-f_t^2)$ (where b is the expected number of mismatches with randomised sequences, and f_x is the proportion of nucleotide x). If base composition varies between two sequences i and j then for very large distances d_{obs} will approach $f_{ai}f_{aj} + f_{ci}f_{cj} + f_{gi}f_{gj} + f_{ti}f_{tj}$ and we can replace b with 1 minus this sum of terms (Bulmer 1991a). If we replace b in the previous equation with this last value then for both zero and "infinite" distances this modified formula

will be exactly correct, and hopefully more accurate than the unmodified distance for all values in between. It is less clear what the coefficient at the front of this equation should be in order to (a) give the best estimate of the unweighted number of substitutions per site or (b) the best distance for reconstructing evolutionary trees (note that different coefficients may optimise for (a) vs (b)). For the present it is arbitrarily taken to be analogue of the stationary case (i.e. -b).

Despite the informal derivation of the "infinite distance" modification to the Jukes-Cantor type equation, it does appear to give a degree of robustness to the analogous distance in the 2-state case (table 3.2) (and in preliminary 4-state examinations, data not shown). What is also interesting is that standard distance distances, without the infinite distance modification, have been extended to accommodate a distribution of rates across sites (e.g. Golding 1983, Olsen 1987, Jin and Nei 1990), are known to have relatively low sampling variances (e.g. Rodriguez *et al.* 1990), and have recently been extended (in the case of the Jukes-Cantor and Kimura 2ST) to allow "unbiased" estimates (Tajima 1993, more on this in appendix 4.2). It will be interesting to see if it is possible to marry together the unbiased and robust part to give a "best of all worlds" distance i.e. robustness plus low variance about the true value, and then extend such a distance to allow a distribution of rates across sites. If these hopes are realised, then this type of approximate distance could be a useful alternative to the implementations of the LogDet (which may suffer problems of sampling variance and sensitivity to unequal rates across sites).

Another point not yet addressed is how best to estimate the relative frequencies of the nucleotides when making an "infinite distance" modification (rather the average base composition of all sites has been used). One obvious solution is to use just the sites that have varied between the two sequences. Using just these sites will negate the "drag" on overall change of base frequency due to sites which have not changed. Sites may not have changed due either to chance under an i.r. model, or to slow or zero rate of change when there is variability of rates across sites (e.g. due to stabilizing selection). When estimating short distances, use of only the varied sites to infer expected base composition could carry a high penalty in increased variance with little gain, as the percentage error due to base composition drift would be low. The biggest gain would be expected at larger distances, and at these distances many sites should be varying. These factors together suggest that a simple weighting function for the proportion of sites to include in estimating base frequency composition of two sequences might be y × (base composition in variable sites) + (1-y) × (base composition in unvaried but potentially variable sites) + 0 × (base composition in sites most unlikely to change). Here the value for y could be an observed (or preferably a reliable inferred) distance between the two sequences being compared (the latter necessitating a new weighting if distance of more than 1 occurred). In the analyses latter in this chapter, the importance of which sites to include when measuring base composition, or performing transformations, is demonstrated.

Going further we can extend the concept of the "infinite distance" correction beyond the Jukes-Cantor type model, to any distance which takes the logarithm or power of a value it expects to go to zero with increasingly large "true" distance. Two obvious examples are the

Kimura 2ST and 3ST distances. The Kimura 2ST distance $\delta = -1/2 \ln [(1 - 2P - Q) (1 - 2Q)^{1/2}]$, becomes,

$$\delta = -1/2 \ln [(1 - 0.5AP - 0.5BQ) (1 - BQ)^{1/2}], \quad (3.3.2\text{-}1)$$

where P = the observed proportion of transitions, A = 1 / (asymptotic frequency of transversions) = $1 / (f_{ai}f_{gj} + f_{gi}f_{aj} + f_{ci}f_{tj} + f_{ti}f_{cj})$, Q = the observed proportion of transversions, while B = $1 / (f_{ai}f_{cj} + f_{ci}f_{aj} + f_{ai}f_{tj} + f_{ti}f_{aj} + f_{ci}f_{gj} + f_{gi}f_{cj} + f_{gi}f_{tj} + f_{ti}f_{gj})$. Similarly the Kimura 2ST distance with a gamma distribution of rates across sites (Golding 1983, Jin and Nei 1990) becomes,

$$\delta = k / 2 \ln [(1 - 0.5AP - 0.5BQ)^{-1/k} + 0.5(1 - BQ)^{-1/k} - 3/2]. \quad (3.3.2\text{-}2)$$

Likewise the Kimura 3ST distance becomes,

$$\delta = -1/8 [M^{-1}(1 - 0.5XP - 0.5YQ) + M^{-1}(1 - 0.5XP - 0.5\,ZR) + M^{-1}(1 - 0.5YQ - 0.5ZR)] \quad (3.3.2\text{-}3)$$

where M^{-1} is the inverse of the moment generating function of the distribution of rates across sites (for a list of closed form formulae see table 2.2), P, Q, and R are the observed proportion of transitions, type 1 transversions and type 2 transversions, while X, Y, Z are the inverses of the expected number of such observed events when the distance has tended to infinity. Tamura (1992) has suggested similar "infinite distance" modifications to an i.r. distance he derives (itself a generalisation of the Kimura 2P method when $f_a = f_t$, $f_c = f_g$, and $f_a \neq f_c$), but gives no discussion of the modification.

Such modifications can also be invoked for more general distances. One example is the general distance solution for a stationary time reversible mechanism as described above (section 3.2.2). Under such a model we expect all **F** matrices to be symmetric. If we observe significant departures from symmetry, we may still invoked the "mechanics" of the distance formula but apply it directly the nonsymmetrised **F** matrix. Doing so will hopefully make the distance estimate more robust in the given application, but there is no guarantee of this. A counter example to the idea that these "infinite distance" modifications can only help is furnished by the Kimura 3ST model. This distance is exact if the substitution process follows the generalised Kimura 3ST matrix, but the root distribution is not at equilibrium (Székely *et al.* 1993, theorem 10), allowing relative base composition to potentially vary substantially between taxa. If we were to apply the "infinite distance" modification to the standard i.r. K 3ST formula, then under such a model the modified distance is not guaranteed to estimate additive evolutionary distances. Before any strong advocating of the general use of the "infinite distance" modifications can be made, it will be necessary to gauge how much more robust they are than the standard formulae. This will necessitate extensive simulations under a spectrum of realistic models of evolution so that their advantages (sometimes increased robustness) and disadvantages (possibly increased variance) can be appreciated. And, of course, it will be important to evaluate whether this class of methods offer any specific advantages over LogDet methods, which otherwise remain the preferred option.

3.4 CONSISTENCY AND ROBUSTNESS UNDER A NON-STATIONARY MODEL

Here a simple non-stationary stochastic model of evolution is constructed, using the equations described in section 3.2. It is shown that three of these six distances often considered to be robust under unequal base composition, can result in inconsistent tree estimation. The first transformed distance (c) considered later in table 3.1, is the 2-state analogue of the Jukes-Cantor, the 2-state equi-frequency Poisson distance (sometimes called the Cavender distance). The next transformed distance (d) is the "infinite distance" modification applied to the unequal base frequency Cavender (1978) model (this is named the modified Cavender distance, which is a 2-state analogue of the Olsen 1987 distance). A third transformed distance (e) is the basic LogDet distance of Steel (1994a) (equation 3.3.1-1), while the fourth (f) is the LogDet - r*ln*(r) formula of Lockhart *et al.* (1994)(where *r* is the number of character states). The fifth estimator (g) is the asynchronous distance proposed by Barry and Hartigan (1987b), and claimed to be consistent under any i.r. / i.i.d. model. Last, (h) is the 2-state analogue of equation 3.3.1-2, the ultimate form of the LogDet of Lockhart *et al.* (1994), which is identical to the Paralinear distance of Lake (1994) (but for the latter lacking an overall coefficient of 1/r)(note this distance is also equal to the averaged asynchronous distances of Barry and Hartigan 1987b, see appendix 3.2).

3.4.1 A model of non-stationary evolution

We use a 2-state (e.g. purines / pyrimidines) Markov model for simplicity, but the fundamental results extend to any number of states. The root distribution is equi-frequency (50:50), and an equi-frequency transition process applies on both internal edges and on the external edges leading to taxa A and B. In contrast on the other two external edges leading to taxa C and D, there has been a change in the substitution process, with a tendency to change state x to y more readily than to change to state y to x. All the external edges have an expected number of 0.2 substitutions per site along them (this number can be verified by applying equation 3.2-1 above to the data given in figure 3.1). We call this a quasi-clock model, to distinguish it from the general molecular clock hypothesis, because although the resultant total amount of evolution is equal in all lineages: (1) the rate of substitution between different states is not constant in time, or across the tree. (2) The total amount of substitution per unit time is generally not equal, but changes as the frequency of the two states change. (Note; if quasi-clocks or something close to them actually exist in biology, as suggested by figure 2.12, one explanation is that a group of organisms can tolerate a certain degree of replication error (which may be selectively elevated in the case of parasites evading host defenses), but the relative proportions of the different types of substitution error change with the replication and repair mechanisms evolution). To assess the consistency of a distance transformation plus tree selection criterion pairing we need to calculate the expected divergence matrix between each pair of species, and we do this using equations already described in section 3.2, and as illustrated in figure 3.1. It is important to appreciate that ultimately all sequences are expected to show some drift in base composition; as with any evolutionary trait there is no known biological mechanism to ensure they stay at any particular equilibrium.

Table 3.1 shows the results of applying tree selection to the data generated in figure 3.1 (in section 3.2) after transformation by the six distance methods being evaluated. The first set of distances in the upper left of the table are the true distances, measured as the average number of substitutions (of all types) per site. Next to the distances are the results of applying a tree selection method to them. Following Buneman's (1971) four point metric, the favoured tree is that with the smallest sum of distances $\delta_{ij} + \delta_{kl}$ (where i, j, k, l refer uniquely to one of the four taxa). With four taxa, this tree selection method is equivalent to neighbor joining or unweighted least squares (tree selection from distance matrices is discussed further in chapter 5). The column marked "$\delta_{ij} + \delta_{kl}$" in table 3.1, gives the sum of pairwise distances $\delta_{ij} + \delta_{kl}$, while an asterisk marks the minimal sum, thus indicating for example, that T_{AC} (A and C together) is the tree selected from the true distances (and indeed this is the tree that generated the data).

We will describe the margin by which the true tree (T_{AC}) is "better" than the incorrect tree T_{AB} by $(\delta_{AC} + \delta_{BD} - \delta_{AB} - \delta_{CD})/(\delta_{AC} + \delta_{BD})$ expressed as a percentage. For the true distances the correct tree has a margin of +7.6%. The observed distances are shown to the right of the true distances in table 3.1. Notice that the most underestimated distance is that from C to D, (the two sequences experiencing non-Poisson evolution), and when basing tree selection on the observed distances, the true tree losses by 5.9%. We are clearly in a zone of inconsistency, which we dub the "Lockhart zone" (as it was Lockhart *et al.* 1992, who with a model similar to the one in figure 3.1 demonstrated that inconsistency of tree selection can occur when base composition varies during evolution).

Table 3.1 Distance matrices and favoured trees for the model in figure 3.1

(a) true distances					$\delta_{ij} + \delta_{kl}$	(b) obs distances				$\delta_{ij} + \delta_{kl}$
	B	C	D			B	C	D		
A	0.431	0.400	0.431	T_{AB}	0.8609	0.289	0.292	0.305	$*T_{AB}$	0.5502*
B		0.431	0.400	$*T_{AC}$	0.8000*		0.305	0.292	T_{AC}	0.5846
C			0.431	T_{AD}	0.8609			0.261	T_{AD}	0.6092
(c) Poisson equi-frequency dist.						(d) infinite Cavender dist.				
	0.431	0.439	0.470	$*T_{AB}$	0.8008*	0.431	0.439	0.470	$*T_{AB}$	0.8263*
		0.470	0.439	T_{AC}	0.8785		0.470	0.439	T_{AC}	0.8785
			0.370	T_{AD}	0.9394			0.395	T_{AD}	0.9394
(e) LogDet F_{ij}						(f) LogDet - rln(r)				
	2.248	2.265	2.326	T_{AB}	4.6515	0.431	0.439	0.470	T_{AB}	0.9394
		2.326	2.265	$*T_{AC}$	4.5296*		0.470	0.439	$*T_{AC}$	0.8785*
			2.403	T_{AD}	4.6515			0.508	T_{AD}	0.9394
(g) B + H distances						(h) LogDet using obs. Π				
	0.431	0.439	0.470	$*T_{AB}$	0.8780*	0.431	0.409	0.439	T_{AB}	0.8780
0.431		0.470	0.439	T_{AC}	0.8785		0.439	0.409	$*T_{AC}$	0.8171*
0.378	0.408		0.447	T_{AD}	0.9394			0.447	T_{AD}	0.8780
0.408	0.378	0.447								

Upper triangles: eight different distances, for sequences from the model given earlier in figure 3.1 (with generating tree T_{AC}). The true distances are measured in number of substitutions per site. All distances except the Barry and Hartigan "asynchronous distances" are symmetric (i.e. $\delta_{ij} = \delta_{ji}$), so it is only for this last distance transformation that we have shown also the lower triangle of pairwise distances estimates. The optimal tree chosen from each upper triangle of distances is indicated by an asterisk.

3.4.2 Inconsistency when using the Barry and Hartigan asynchronous distance

We now consider how tree selection fares after making our six different distance transformations to the observed data. The 2-state equifrequency Poisson distance (c), is inconsistent, selecting tree T_{AB} with a margin of 8.8% over the true tree (T_{AC}, table 3.1). This is an even worse inconsistency than with the observed distances. Interestingly this increased degree of inconsistency comes about because this transformation results in overestimating the distance between sequences with quite unlike base compositions i.e. δ_{AC} and δ_{BD} (and not due solely to an underestimation of the distance between sequences C and D). In this example the modified "infinite distance" Cavender distance (d) transformation makes a better estimate of the distance between the sequences C and D, and for this reason suffers less severe inconsistency (margin 5.9%). The "infinite distance" modification has resulted in improved robustness, but tree selection is still clearly inconsistent.

Next we consider how the straight LogDet distance (e) of Steel (1994a) performs. This method is perfectly tree additive, as can be seen by the two incorrect trees having the same score. The margin of tree T_{AC} over either incorrect tree is a reduced 2.7%, due to this LogDet measure effectively having a large constant added to all distances (this constant being LogDet of Π_R - notice the conspicuously large size of these distances in table 3.1). The LogDet -rln(r) distance

(f) is also perfectly tree additive, while the constant +rln(r) has resulted in the margin by which the true tree is favoured increasing to 6.9%. The Barry and Hartigan (1987b) form of LogDet, their "asynchronous distance" (g), attempts to remove the effect of the root distribution of states by subtracting the log of the determinant of the matrix of state frequencies in the first species of a pair (e.g. Π_A). Despite Barry and Hartigan's (1987b) claim, proposed as a mathematical proof, this is not a tree additive distance under a non-stationary model and so can lead to inconsistency of tree selection. Table 3.1 shows that if we take our model sequences in the order A, B, C, D, then use the Barry and Hartigan (1987b) "asynchronous distance" measure, our tree selection method chooses T_{AB}, an incorrect tree (the margin here is just 0.1% as we selected this example to be right on the point at which inconsistency occurs).

The reason for the non-additivity of Barry and Hartigan's "asynchronous distance" is that the amount subtracted from the tree additive measure LogDet **F** (since $\ln[\det(\mathbf{P}_{mn})] = \ln[\det(\mathbf{F}_{mn})] - \ln[\det(\Pi_m)]$) is conditional on the labeling of the sequences. This non-additivity leads to inconsistency of tree selection with certain tree shapes, and transition matrices along edges. The reason for the incorrect tree selection when using the "asynchronous distance" in table 3.1, is that the Barry and Hartigan distance has overestimated all the distances between sequences with differing base composition, while those distances measured between sequences of equal base composition are correct (i.e. they are the same as the additive distances estimated by equation 3.3.1-2). Notice that if we took the sequences in reverse order (the lower half of the distance matrix in table 3.1) then we would choose the right tree. However, even here we have undesirable effects, giving too much weight to the difference between trees, which has the effect of generating false confidence in the optimal tree when we apply a statistical procedure (e.g. the bootstrap) to this data.

An irony of the history of these distances is that Barry and Hartigan (1987b) when faced with a situation of unequal base composition amongst species (in their illustrative example), resorted to averaging their distance. Unfortunately they did not justify this averaging and throughout their paper emphasized the consistency of their asymmetric, "asynchronous distance" which they wrongly believed to be additive under the conditions of general i.i.d. and i.r. model (as their abstract clearly states). As we have shown in appendix 3.1, $-1/8[\text{LogDet}(\mathbf{P}_{mn}) + \text{LogDet}(\mathbf{P}_{nm})]$ (the average of the "asynchronous distance pairs) is another form of the LogDet equation $\delta = -1/4p[\text{LogDet}(\mathbf{F}_{mn}) - 1/2\ (\text{LogDet}(\Pi_m) + \text{LogDet}(\Pi_n)]$ (this identity became apparent from discussions with John Hartigan in August 1994). Lastly when the ultimate form of LogDet (3.3.1-2) is used to transform our model data, tree selection chooses the true tree T_{AC} with a margin 7.5% over the incorrect trees (table 3.1). Because the two non-optimal trees have an equal sum of $\delta_{ij} + \delta_{kl}$, we are assured that in this example (and as proven generally by Steel 1994a, Steel *et al.* 1993c, and Lake 1994) that these distances are tree additive.

Another interesting feature shown in table 3.1 is that when the base composition varies between species, then the LogDet measure Lockhart *et al.* (1994)given in equation 3.3.1-2, returns more reliable estimates of the average number of substitutions per site than any of the other distance measures. This occurs even though the weighting of substitutions moves away

from unity as the sequences evolve to have non-equifrequency base composition. As table 3.1 shows, when the LogDet distance is measured upon sequences evolving via a non-stationary i.i.d. model, it can return values which are both greater than or less than the "true" distance measured as total number of substitutions per site. This suggests that if a biologist is interested in edge lengths, yet the data are clearly from a non-stationary model, then he should be prepared to consider that a tree reconstructed from LogDet distances may have more reliable edge length estimates than a tree constructed from any of the other presently available distances. This insight should be useful when ball park divergence date calibrations are to be made with non-stationary data. Further, given a reasonable estimate of the base composition at the internal nodes of a tree and given our interpretation (3.3.1-6) of what the LogDet distance (3.3.1-2) is measuring, then it is possible to infer whether this distance is over or underestimating the total number of substitutions per site on a particular edge. A similar inference can be made for other pairwise distance corrections, but their non-additivity under nonstationary models makes the interpretation of edge lengths on a tree reconstructed from many pairwise distances much more difficult to interpret.

Surprisingly, the LogDet in the form of equation 3.3.1-2 not only helps in tree estimation, but under conditions of nonstationary base composition, the edge lengths on a tree built from these distances may be more realistic than any other. As mentioned already in earlier sections, under stationary conditions with identical rates across sites and long sequences, LogDet distances in combination with a statistically efficient tree reconstruction method are expected to give the most reliable relative edge lengths of any distance method. These findings are encouraging for the general use of LogDet distances in molecular phylogenetics. They suggest, for example, that LogDet distances will be favoured for estimating the relative divergence times of taxa, using long non-coding sequences (e.g. estimating the divergence times of mammalian orders, given sequences and fossil calibrations times in the more recent Tertiary).

3.4.3 Oh No! Long edges can repel

We now consider permutations of the model in figure 3.2, and describe the robustness of the different distance transformations under these related models. Different models were generated by simply permuting the transition matrices of the external edges (of the tree in fig. 3.1), and also in one case transposing one of these transition matrices (which is equivalent to evolution with the opposite trend to the untransposed matrix, e.g instead of a sequence becoming GC rich it becomes AT rich). These models are illustrated in figure 3.2 (below). The results of tree selection-distance transformation combinations applied to the data from these models, are given in table 3.2. The robustness of the different distance transformations under the four models of figure 3.2 are shown in table 3.2. Table 3.2 also adds the minimum length (X) that the internal edge on the tree in figure 3.1 must be in order to maintain consistency when using a specific transformation. Notice that under model M1, the value of X is smallest for the Barry and Hartigan "asynchronous distances", which is in agreement with the already described "margin" statistic to measure preference for T_{AB} over T_{AC}. Notice that the "infinite distance" modification

to the Cavender distance has slightly improved robustness, relative to the unmodified Poisson correction.

Model M2 is constructed by simply reversing the non-stationary process on one of the external edges in model M1 (see figure 3.2). Under M2 the "asynchronous distance" is just as susceptible to inconsistency as under M1. This is because the correction factor which needs to be added to the "asynchronous distance" $\ln[\det(P_{mn})]$ in order to make it identical to the additive LogDet formula in equation (3.3.1-2) has the same value in both cases (specifically the factor to add to make it tree additive is $(1/2\ln[\det(\Pi_m)]-1/2\ln[\det(\Pi_n)])$ and here since Π_n of M2 is just Π_n of M1 with the order of the diagonal elements reversed, then $\det \Pi_n$ is equal in both cases). Note that under M2, tree selection based upon either the Poisson or the modified Cavender distance is consistent. This is because the biggest error occurring within this model, is that these last two distance estimates very much overestimate the true distance from sequence C to D. For this model it ensures consistency, but we will see in M4 that this same factor results in inconsistency. Because this is a systematic error it is not desirable, and can play havoc with estimating the statistical support for a particular tree (which as we have discussed in chapter 1 is of prime importance).

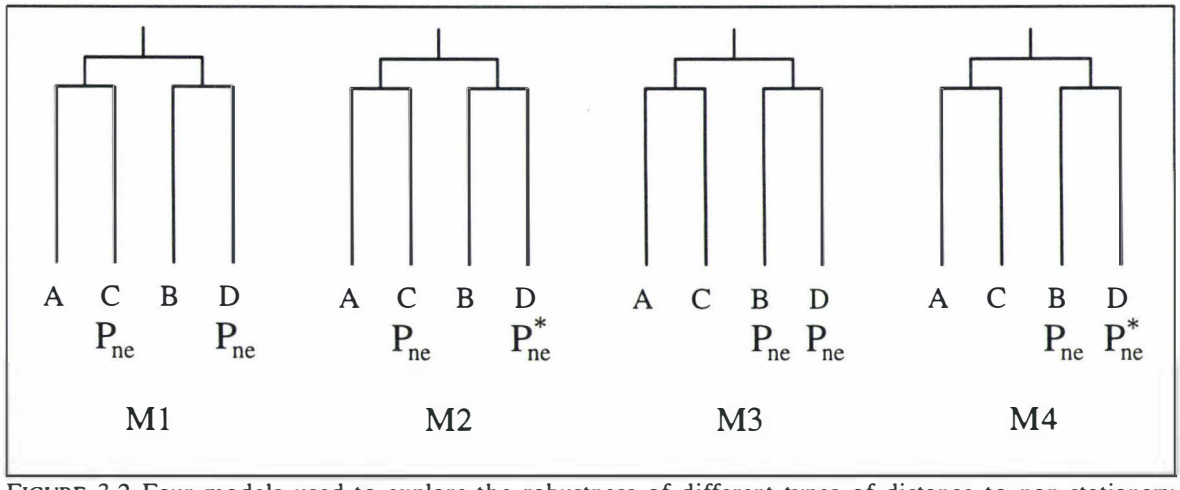


FIGURE 3.2 Four models used to explore the robustness of different types of distance to non-stationary Markov processes on a tree. M1 is the model described in figure 3.1, where $\mathbf{P}_{ne}$ is the non equifrequency transition matrix [0.98 0.02, 0.36 0.64] (where the comma denotes the end of the first row). The superscript * denotes the reverse of this matrix namely [0.64 0.36, 0.02 0.98], which is a non equifrequency Markov process going in exactly the opposite direction to $\mathbf{P}_{ne}$. Note this reverse matrix is not the transpose.

Table 3.2 Consistency of tree selection upon the models described in figure 3.2

	M1		M2		M3		M4	
distance	X	opt. tree	X	opt. tree	X	opt. tree	X	opt. tree
B + H	0.030	T_{AB}	0.030	T_{AB}	consistent	T_{AC}	consistent	T_{AC}
Cav. mod.	0.058	T_{AB}	consistent	T_{AC}	consistent	T_{AC}	0.058	T_{AB}, T_{AD}
Poisson	0.075	T_{AB}	consistent	T_{AC}	consistent	T_{AC}	0.089	T_{AB}, T_{AD}

Consistency of different distance corrections combined with Buneman's four point metric for selecting the correct unweighted tree for the models (M1-M4) described in figures 3.1 and 3.2. The value X is the length of the internal edge of the unrooted tree (measured as expected number of substitutions per site) below which inconsistency occurs, so the larger X the worse the inconsistency problem. The column "opt. tree" indicates which tree(s) is optimal (that is the tree selected) as X drops into the range at which inconsistency occurs. In all cases the correct tree is T_{AC}. Distance values are taken from the upper right of a table of distances (important in the case of the Barry and Hartigan distance, since it is often asymmetric with these models). If we were to take the lower left "asynchronous distances", they are consistent on M1, and M2, but not on M3 and M4 where inconsistency occurs with the tree T_{AB} becoming optimal for small X. If the "asynchronous distances" are averaged, this generates the LogDet distance which is consistent in all cases.

Now two other possible permutations of model M1, are examined. The third model M3, has sequences with like base composition grouped together on the same tree, and in this case tree selection can't help but put sequences with like base composition together. Model M4 however shows an interesting new feature "unequal base compositions repelling." Here we see the Poisson and the modified Cavender distances greatly overestimating the distance between the two sequences B and D which have unequal base composition (the actual distance should be 0.4, the estimated distance was 0.657 {Poisson} and 0.594 {mod. Cavender}). Indeed the Poisson and modified Cavender distances overestimated all true distances between sequences with unequal base composition, with the worst overestimation being between those sequences with the most unequal base composition. Under model M4, with the internal edge length set at 0.09, the true distances were 0.49(δ_{AB}), 0.40(δ_{AC}), 0.49(δ_{AD}), 0.49(δ_{BC}), 0.49(δ_{BD}), 0.49(δ_{CD}), the observed distances were 0.326, 0.276, 0.326, 0.326, 0.366, 0.326, the Poisson corrected distances were 0.529, 0.4, 0.529, 0.529, 0.657, 0.529, and the modified Cavender distances were 0.529, 0.4, 0.529, 0.529, 0.594, 0.529 respectively. This "over-estimation" effect was enough to cause either tree that separates B and D, to be optimal.

This result, that standard distance estimators can dramatically overestimate the true distance, is new and emphasizes an unappreciated aspect of unequal base composition effects. Contrary to the emphasis in Lockhart *et al.* (1994), in this example the two sequences with the most similar base composition are not grouping together, but rather unlike sequences are being repelled apart by standard estimators. Thus it seems reasonable to say that "edges with unequal base compositions repel", as well as "sequences with similar base composition attract", at least for tree reconstruction using pairwise distances and the current selection of standard estimators. It also follows that generally the longer two edges are, the more opportunity there has been for base composition to move further in opposite directions, hence it is sometimes appropriate to say "long edges can repel" as well as attract. In the second half of this chapter we encounter a

situation where unequal base compositions repelling, adds to the difficulty of resolving the earliest branching event amongst eukaryotes. The observed distances were all clearly underestimates, but there was no "anti-Lockhart" effect acting, these distances were somewhat more robust than the Poisson distances e.g. with model M4, X had to be less than 0.74 before inconsistency of tree selection occurred.

These findings thus refine the understanding of Lockhart *et al.* (1994) of how base composition can affect tree reconstruction. They also give a valuable example for phylogenetics. It seems that there is a tacit assumption in phylogenetics "that a method which corrects for some multiple hits must be better than no correction." Clearly this need not be the case; prior to these analyses it was expected that the Poisson transform would be only partly accounting for multiple hits, because the actual model was more complex than the model the transformation is based on. That this is not the case raises the question of how as yet unappreciated artifacts of any of our current "transforms" could result in incorrect trees, especially when analysing anciently diverged functional molecules. Another tacit assumption which may also need reconsideration is that it is all right to construct methods piecemeal, "i.e. if a helps and b helps, then a method with a plus b, will help more." Examples of this sort of method abound (e.g. combinations of different sorts of "weighting" in parsimony, distance or ML analyses). The only solution is to study how our current methods behave under more general models which show all the major features of molecular evolution (some of which we are only beginning to understand). As an example of the unusual results that can arise, consider applying standard logarithmic transformations to four sequences, where two sequences have both equal and the median base composition. In this situation the favoured trees due to unequal base composition induced systematic error are either of the trees that separate the two sequences with unlike base composition, and not the tree that puts the two sequences with identical base composition together. The general problem of what happens when some distances are systematically overestimated is addressed in chapter 5 under the title of the "anti-Felsenstein zone" problem.

In conclusion then, results of these analyses show that the Barry and Hartigan "asynchronous distance" LogDet($\mathbf{P}_{mn}$) can lead to inconsistent tree estimation, when the Cavender distances are consistent, and vice-versa. In contrast the LogDet($\mathbf{F}$) methods of Steel (1994a), Lockhart *et al.* (1994), and Lake (1994) are always tree additive for any i.i.d. and i.r. Markov model (proven in the aforementioned references). Interestingly the infinite distance modification to the Cavender distance made it a bit more robust than the equi-frequency Cavender distance, suggesting it may be a useful modification in some circumstances. The coefficient 1/b (= 1 - expected match frequency) in front of this distance also seemed appropriate, as it was making the distance closer to its true value than the factor 1/2. For all these models (M1 to M4), the LogDet distance $\delta_{mn} = -1/4(\text{LogDet}(\mathbf{F}_{mn}) - 1/2\ \text{LogDet}(\Pi_m) - 1/2\ \text{LogDet}(\Pi_n))$, was the most reliable indicator of the "true" distance (substitutions per site) amongst those compared. These analyses have also turned up unexpected results such as the strong repulsion that sequences with unequal base compositions can have (especially when transformed with standard methods). Tacitly this gives encouragement to the use of "infinite distance" modified standard distances, or better still,

LogDet transforms. Importantly, it also reminds us that there are big gaps in our understanding of how systematic errors can occur. One more piece of this puzzle is filled in chapter 5, where the "anti-Lockhart zone" effect is put in a more general context of "long edges repel" (which is dubbed the "anti-Felsenstein zone" problem).

3.4.4 A brief history of LogDet distances

No fewer than five research groups independently developed and appreciated in various degrees the importance of LogDet type transformations to genetic distance estimation. These are the publications on the method that we are aware of.

There have been various specific uses of LogDet type equations in the area of Markov modeling going back decades. We do not review these, but do note that an application of LogDet equations which is essentially like that of phylogenetics is given in Pearl and Tarsi (1986). In 1986 Rodríguez and Medina (unpublished) developed the distanced δ_{ij} = -1/4 (ln[det(**F**)]-ln[det(Π)], where Π is nucleotide frequency at the root) under a stationary molecular clock model. They noted that this distance would be additive for any i.r. and i.i.d. model, and noted that δ_{ij} is a summation of the diagonal terms of the instantaneous rate matrix given a stationary base composition, for example, det(**P**) = det(**P**^{t}) = det(exp(**R**t)) = exp(tr(**R**t)), Bellman (1970). This manuscript also developed a solution to the general i.i.d. and i.r. reversible rate model (after rejection of this first manuscript only this latter part was published in Rodríguez *et al.* 1990). Another paper of the same period, Cavender and Felsenstein (1987) described the log of the determinant of internal edges of a tree as a measure of edge length and showed that it is tree additive, but did not use it for any specific purpose.

Barry and Hartigan (1987b) described what they called an "asynchronous distance", δ_{mn} = -1/4ln[det(**P**$_{mn}$)]. Despite assuming that this distance was tree additive, and so consistent, under non-stationary models (shown to be incorrect in sections 3.4.1 and 3.4.2) they produced many useful results. They gave a useful interpretation of their distance meant in terms of weighted nucleotide substitutions (which again can be thought of as the trace of **R**; we used this results earlier in 3.3.1-6 to infer what the tree additive LogDet distance of Lockhart *et al.* (1994) and Lake (1994) is measuring). They also derived a delta method approximation to the variance of LogDet(**P**), (which Lockhart *et al.* 1994 then used to derive the variance of LogDet(**F**), 3.3.1-1), and studied how adequately this estimated variance could be used to construct a confidence interval about δ. Their paper also contained the same results for the time reversible distance correction formulae (definition, proofs, a variance estimate, and evaluation of confidence intervals). At the end of their paper they analysed some primate mtDNA sequences using these distance estimates, averaging the asynchronous distances (although they did not justify this, nor did they advocate it). At this point they missed the opportunity to reexamine their results and realise that it is the averaged asynchronous distance only which is tree additive under non-stationary models.

Chang and Hartigan (1991) looked again at asynchronous distances, but again they were in error claiming that asynchronous distances are additive under the general i.r. and i.i.d. model. Unfortunately this error undermines their claims of being able to identify the generating tree under any i.r and i.i.d. model.

After a delay of lying dormant, two researchers (M.A. Steel and J.A. Lake) independently realised that the log of the determinant of the matrix **F** would be a tree additive distance under the general i.r. and i.i.d. model. Steel (1994a) constructs a proof a mathematical proof of these properties for any number of states, describing exactly what conditions must be met for the distance measure to be tree additive. He also used the result in reverse to show that all i.i.d. and i.r. evolutionary models will produce a unique set of sequence pattern probabilities, and that any tree of sequences evolving under this model could be recovered by use of the LogDet transformation, and a consistent tree building method (or by the general 12 parameter per edge ML method of Barry and Hartigan 1987a). (Pers. comm. from Terry Speed notes that Trang Nguyen independently used this approach to show the uniqueness of sequence patterns in her 1991 Ph.D. thesis at UC Berkeley). Dr Steel's discovery of the LogDet method was prompted by Dr P.J. Lockhart's ongoing search to find a method that could better cope with unequal base composition effects (e.g. Lockhart *et al.* 1992). This resulted in the paper Lockhart *et al.* (1994), which also specified various ways to subtract the influence of the ancestral base composition from the distance in Steel (1994a). Most importantly, they emphasized with real examples the important role that LogDet distances can play by apparently overcoming unequal base composition effects in real sequences, ensuring that biologists would appreciate the advantage that these distances potentially offer. Lockhart *et al.* (1994) voiced a note of caution that the LogDet was not robust to unequal substitution rates between sites. In their study they suggested that reducing sequences to just those sites which were parsimony informative, then applying the LogDet transform gives more reliable results than using all sites, but did not explain why.

J.A. Lake described his findings in UCLA patent case LA 91-035-01 (May 31, 1991) and Lake (1994)(the patent application was withdrawn in 1994). Lake (1994) derives the most refined version of the LogDet distance, calling it a "paralinear" distance. Using a different approach to previous authors, he constructed a partial mathematical proof of the asymptotic ($c \rightarrow \infty$)additivity of this distance estimate (his proof is not unambiguous, M.A. Steel pers comm.). He also emphasised that the this distance is sensitive to unequal rates across sites, but from a numerical example suggested it is intrinsically more robust than traditional distance measures. Below we show that this extra robustness is not generally true, but then go on to show that there are modifications that can give both LogDet and other i.r. distance methods a great deal more robustness when rates at sites differ.

3.5 MAKING LOGDET DISTANCES ROBUST TO RATES ACROSS SITES

We now look at a modification to distance estimates which gives them much enhanced robustness to unequal rates across sites. Unfortunately the apparently sensible thing step of replacing the ln(det(**F**)) distance with an M^{-1}(det(**F**)) when there is a distribution of rates across sites does not work. The reason for this is that det(**A** + **B**) ≠ det(**A**) + det(**B**), where **A** is $\mathbf{F}_{mn}$ at rate λ, and **B** is $\mathbf{F}_{mn}$ for rate xλ, with sites otherwise having the same instantaneous relative rates of substitution. One promising exception to this apparent bind is the invariant sites model which we now examine in detail in the hope that it in combination with the LogDet will produce possibly the most robust method of transforming sequences into tree additive pairwise distances.

3.5.1 Four ways to modify distances to be consistent under invariant sites models

As we have already seen in the case of Hadamard conjugations (and with more relevant results in chapter 5), to a first order of approximation, the invariant sites model compensates for unequal rates at different sites. For this reason we pay special attention to invariant sites / LogDet models, in the hope that removing just the right proportion of unvaried sites will give the method a large degree of robustness under any distribution of rates across sites. This is quite crucial to a distance method and we show (counter to the example in Lake 1994) that LogDet distances are just as prone to error due to unequal rates across sites as are more commonly used distances such as the Jukes-Cantor or Kimura 2P. In this section we detail important steps in how to separate out invariant sites, before going on look in detail at the best way to estimate the proportion of sites to be excluded. The same principals discussed here apply, with a little modification, to other methods which are influenced by the base composition of unvaried sites, for example maximum likelihood models.

For distance estimators which use the relative frequencies of AA, CC, GG, and TT matches and not just their sum (as with the generalised Kimura 3ST method) there are a number of ways of removing a proportion of invariant sites. The model we will assume for this section will be one in which we have invariant sites which do not change identity (i.e. not a covarion model, Fitch and Markowitz 1970) with the remainder of sites being i.i.d and i.r.

Let **D** be the 4 × 4 matrix where entries correspond to the number of times sequence *m* has state *i* and sequence *n* has state *j* in a pair of aligned sites. Under our invariant sites model **D** = **V** + $c\Pi_{inv}$, where **V** is the divergence matrix of the sites able to undergo substitution, *c* is the sequence length, while Π_{inv} is the diagonal matrix of the relative proportions of invariant sites AA, CC, GG, and TT. The unvaried sites are distributed in an unspecified way between the diagonals of **V** and $c\Pi_{inv}$. To remove the effect of the invariant sites we must subtract $c\Pi_{inv}$ from **D** before calculating $\mathbf{F}_{-pinv}$ or applying any correction formula, so let **V** = **D** - $c\Pi_{inv}$. If we wished to express our divergence matrix as proportions (that is dividing all entries in **D** by the sequence length, c) then $\mathbf{F}_{all}$ = **D** / c , where $\mathbf{F}_{all}$ is the normalized divergence matrix of all sites. $\mathbf{F}_{-pinv}$ = (**D** - $c\Pi_{inv}$) / (c [1-p_{inv}]) = (**F** - Π_{inv}) / [1-p_{inv}]. So for our purpose $\mathbf{F}_{-pinv}$ is our estimate of **V** / c.

We now examine a number of ways to estimate Π_{inv} with 4-state sequences; we describe each in turn and briefly describe their inherent assumptions. To illustrate how different ways of estimating Π_{inv} can change the additivity of a transformation of $\mathbf{F}_{-pinv}$ we will plot the sum of squares residual of a tree reconstruction method as we assume different values and forms of Π_{inv}. The tree reconstruction method we are using is weighted least squares (see Felsenstein 1982). The distance estimate is $\delta_{mn} = \frac{-(\ln(\det(\mathbf{F}_{mn}) - r\ln(r))}{r}$, and the weights are the estimated variance of this distance (where r is the number of character states, see Lockhart *et al.* 1994, eq3). The variance of this form of LogDet is equal to $1/r^2$ the variance of $-\ln[\det(\mathbf{F}_{mn})]$ (since the factor $r\ln(r)$ is a constant). An estimate of the variance of $-\ln[\det(\mathbf{F}_{mn})]$ is given by Lockhart *et al.* (1994) equation 2. Rewriting this variance estimate in matrix form and adding in the factor $1/r^2$ we have,

$$\mathrm{Var}[1/4(-\ln[\det(\mathbf{F}_{mn})])] = (\mathrm{trace}(\mathbf{BF}_{mn})-r^2) / 16c, \qquad (3.5.1\text{-}1)$$

where **B** is the matrix whose *ij*-th entry is the square of the *ij*-th entry of $\mathbf{F}^{-1}$ (inverse of $\mathbf{F}_{mn}$). We chose to use this variant of the LogDet because it gives additive distances for an i.i.d. and i.r. model (as we assume all variable sites are thus distributed), and because a closed form variance estimate is currently available. For illustration purpose we will apply our different methods to the 16S-like rRNA sequences used for figure 2.7 and described in chapter 1 (as aligned by Lake 1988).

Each of the methods below requires an estimate of the proportion of sites which are invariant (this overall proportion is called, p_{inv}). We examine different ways of obtaining this estimate later in section 3.6. For now we assume we know p_{inv}, and consider the following ways of estimating what Π_{inv} is.

Method (1). Having been supplied with a value of p_{inv} a straight-forward approach to inferring the diagonal elements of Π_{inv} is to allocate invariant sites equally amongst the character states r, that is,

$$\Pi_{inv}(1) = (p_{inv} / r) \times \mathbf{I}, \qquad (3.5.1\text{-}2)$$

where **I** is the identity matrix. This model assumes automatically that invariant sites are equi-frequent. Despite this being unlikely to be exactly correct, the removal of invariant sites by this method shows dramatic results, making our distance data much more tree additive (figure 3.3).

Method (2). Since nucleotide frequencies are often clearly unequal in real sequences, we may expect there to be a strong correlation between the average proportion of each nucleotide in the sequences (Π) and the entries in Π_{inv}. Accordingly we calculate,

$$\Pi_{inv}(2) = p_{inv} \times \Pi. \qquad (3.5.1\text{-}3)$$

Application of this method to the four sequences gave nearly identical curves to those in figure 3.3, but noticeably the point of best fit moved to a slightly higher value of p_{inv} (about 0.14), while

the steep right hand side of the curves (an apparent asymptote) moved from about 0.23 to 0.25. Both these observations suggest that this method is a more realistic estimation of invariant sites in this data. In particular, we expect the asymptote for this LogDet transformation to occur at smaller values of p_{inv} when more sites are removed from a particular class of unvaried sites than should be (because this drives the determinant of $\mathbf{F}_{\cdot pinv}$ more quickly towards zero).

Method (3). Since we know that sites in $c\mathbf{P}_{inv}$ are amongst the unvaried sites then we may wish to replace the all sites Π with Π_u estimated from just the unvaried (constant) sites, thus

$$\Pi_{inv}(3) = p_{inv} \times \Pi_u, \qquad (3.5.1\text{-}4)$$

For our data this method shifts the point of best fit (and the apparent asymptote) to coincide with slightly higher values of p_{inv} than either of the previous two methods, suggesting that it is a more accurate estimate of the invariant sites. Method (3) will estimate Π_{inv} with higher variance than method (2), so it may be desirable to use this method only when there is good evidence that the frequencies of the invariant sites are different to those of the variable sites, or when there are large amounts of change so that the unvaried sites are dominated by invariant sites. Indeed if our model of strictly invariant and i.r. variable sites holds, then Π_u will converge to Π_{inv} of the invariant sites as new aligned sequences are added (assuming they are different to each other and do not contain sequencing errors). When this method of removing "invariant sites" was applied to the data used in figure 3.3, the inferred number of invariant sites again increased slightly.

Method (4). Given a measure of the fit of our distances to a tree, we numerically optimise the four entries in Π_{inv} (while the sum of the diagonal of this matrix would give p_{inv}). An explicit optimisation must improve the fit of an invariant sites model to the data to model and may also be expected to move the asymptote to higher values of p_{inv}. One fit measure for this purpose would be generalised least squares (see chapter 5). An alternative approach would be to optimise by maximum likelihood (or a related method such as minimum X^2) the parameters of an invariant sites model of the sequence data. For example, if we were using the general LogDet transformation, then the corresponding ML model would be the general 12 parameter per edge (where we directly optimise the values in the transition matrix of each edge with no concern for whether it describes a single rate matrix on that edge) i.i.d. and i.r. model (like that described by Barry and Hartigan 1987a), plus four proportions (three of which are independent) of invariant sites for A, C, G, and T. For a moderate or large number of taxa, the ML method may be preferable as it does not require explicit evaluation and manipulation of a covariance matrix of order t^2. However the optimal Π_{inv} for the ML method may differ slightly from Π_{inv}, with the latter perhaps being the more useful in cases of moderate model violation. Simpler models could also give reliable estimates of Π_{inv} if the data were not especially non-stationary, and the underlying transition process was well approximates by the ML methods assumptions. In some circumstances it might also be important to use other biological information to estimate the proportions of each type of invariant site (e.g. information from other sequences not included in the study).

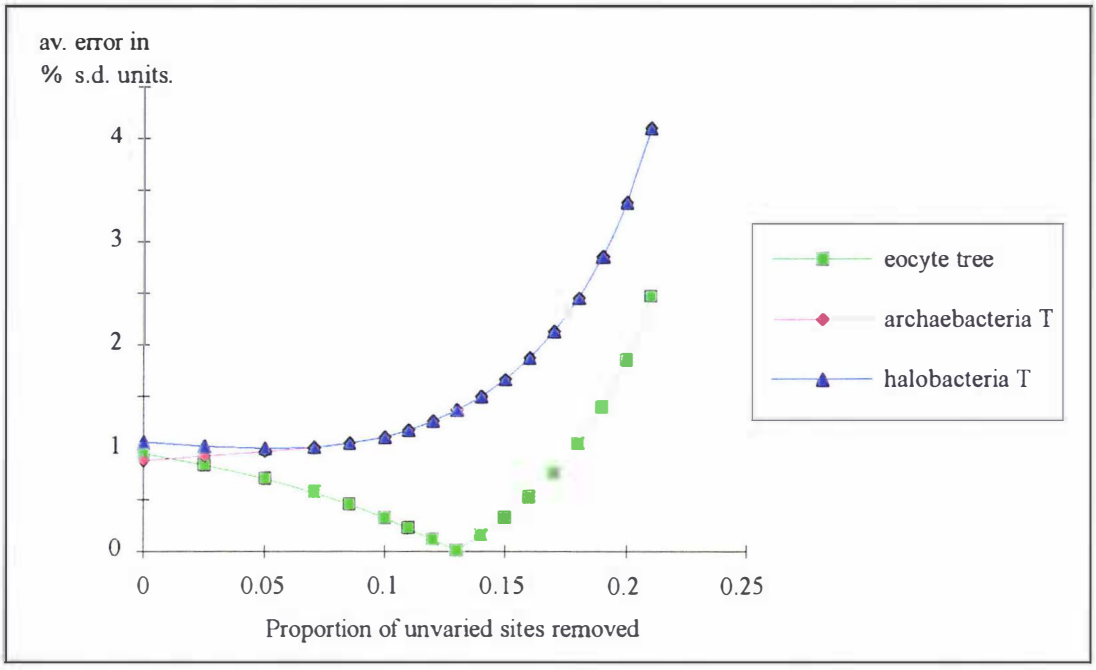


FIGURE 3.3 The weighted least squares fit of the three resolved unrooted trees to the LogDet distances (using the $r\ln(r)$ adjustment factor and minus invariant sites by method (1) from four sequences of Lake (1988)(see section 1.9.1). The measure of fit, is equal to the average amount (measured in percent of a standard deviation) by which observed and tree distances disagree. The halobacteria tree collapses to the star tree (i.e. no support for an internal edge) with no unvaried sites removed, while the archaebacteria tree collapses with > 7% of sites assumed invariant. Interestingly while the model is fully parameterised (i.e. 6 distances and 6 parameters, 5 edge weights plus p_{inv}) we hypothesise that it still has some discriminatory power because of the qualitative aspect of the model, i.e. different unweighted trees. Due to the high standard errors on the estimated distances, these trees are not clearly resolved in this example.

In summary then we have four main ways to apportion invariant sites into the four classes A, C, G, T. Whichever approach(s) is / are used can be checked for appropriateness by the types of studies made in section 3.6 on base composition in sequences. Lastly, by removing unvaried sites directly from the diagonal of each **D** matrix, prior to forming the **F** matrices, it is possible to effectively remove more "unvaried sites" from pairs of taxa than are unvaried across all taxa. This may be appropriate when the model does not hold exactly, yet this is the treatment which yields the best additivity of distances. Such a situation might also arise under a covarion invariant sites model (where invariant sites are only locally invariant), like that discussed in Fitch and Markowitz (1970).

3.5.2 A direct look at the robustness of the invariant sites-LogDet method

As discussed in the previous section, for the most general i.i.d. model it is not yet possible to model a continuous distribution of rates across sites without first separating sites into rate classes (since the determinant of a sum of transition matrices at different rates, is not equal to the sum of determinants of those same matrices). One variable rates at sites model which can be made exact in combination with the LogDet distance transformation is the invariant sites model (Fitch and Markowitz 1970), as described earlier in section 2.3.4. Throughout this thesis we argue that removing a certain proportion of invariant sites, offers a robust first order approximation to any distribution of rates across sites. We now present a further piece of

evidence to support this claim in the form of figure 3.4. This shows how well the invariant sites plus LogDet transformation can estimate an additive distance, when the true model is a Kimura 3P process with a substantial distribution of rates across sites according to a gamma distribution with shape parameter of 1. The parameters we have chosen for this examination are very similar to estimates we obtain (and have obtained, chapter 2) from 16S-like rRNA sequences (which are analysed extensively in the following sections), and the range of distances are also very similar.

Any distance which gives a perfectly straight line with positive slope in figure 3.4 is additive, a crucial property for phylogenetic estimation. As can be seen in this figure, the distance transformations assuming identical rates of change across sites become dramatically non-additive when the true distance (x-axis) exceeds 0.4 substitutions per site. (Also note that interestingly the difference between the 2P and the 3P model is minimal, despite the two fold difference in transversion rates). However, the invariant sites LogDet transformation with either 0.2 or 0.22 removed from the diagonal of $\mathbf{F}$ (in this numerical example methods 1-4 of section 3.5.1 are equivalent), offers a good first order approximation to additivity for the distances considered. Note also that while the LogDet transformation with invariant sites parameter p_{inv} = 0.22 tends to overestimate the shorter distances, it appears overall to generate a straighter line than with p_{inv} = 0.2 or 0.17. An important implication here is that LogDet distance estimates are no more robust to a distribution of rates across sites than are standard distance measures. This finding is contrary to a claim by Lake (1994) of some robustness of tree selection to a distribution of rates across sites when a LogDet transformation was first performed. In fact the robustness gain of the LogDet method over all other methods in his simulations (including the Kimura 2P transformation, table 1, Lake 1994) is best attributed to the model used violating all other methods expectations of the form of the underlying transition rate matrices (e.g. equal rates to tv1 and tv2 transitions), which did not hold.

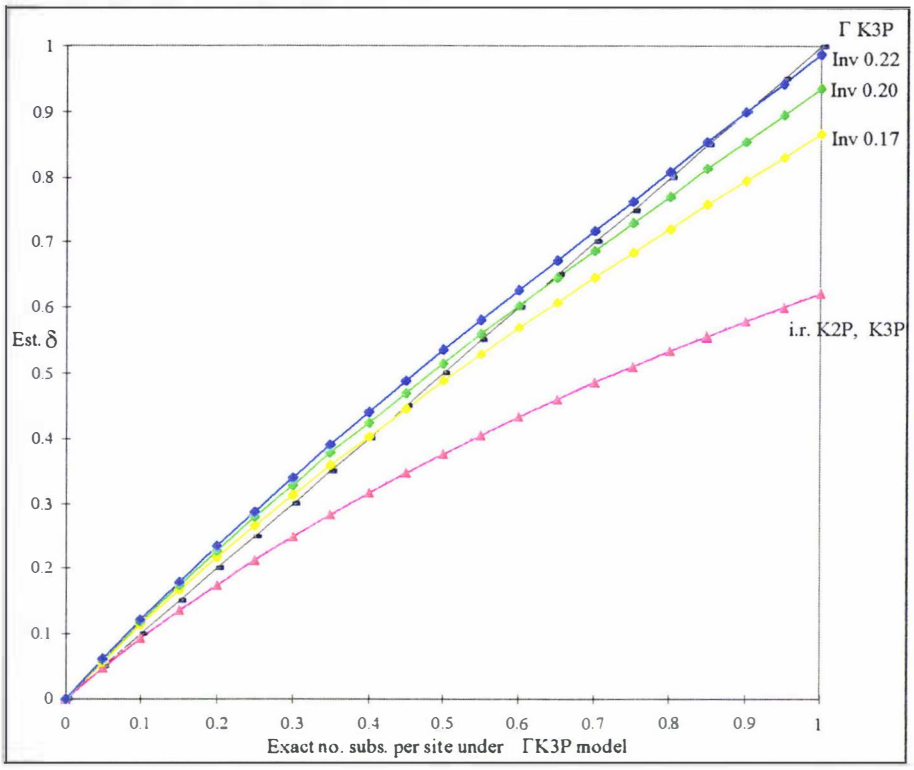


FIGURE 3.4. Approximation of the true distance with a gamma distributed rate model by a LogDet invariant sites model. The observed sequence patterns are calculated under Kimura 3 parameter model (parameters set to tr = 5, tv1 = 2, tv2 = 1) with a gamma (Γ) distribution of rates across sites, k = 1 (i.e. an exponential distribution). An additive distance measured in unweighted expected no. substitutions per site is indicated by the diagonal black line (Γ K3P). The green line marked "Inv 0.2" indicates the number of substitutions estimated by the invariant-sites LogDet transform with 20% of unvaried sites removed.

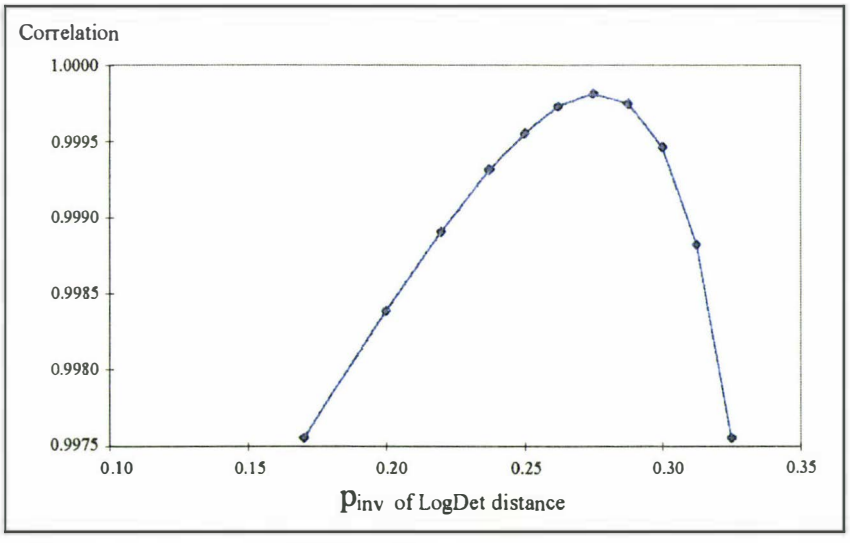


FIGURE 3.5. The linearity (and hence additivity) of the invariant sites - LogDet distance with different values of p_{inv} with respect to the true distance for the data in figure 3.4. Here the linearity is measured by the correlation coefficient with 1 indicating a perfect linear relationship. The optimal linearity occurs at $p_{inv} \approx 0.275$.

Since it is the straightness of the line that measures additivity (for our example) we investigated this feature further. We measured the linearity of the invariant sites LogDet distance

for various values of p_{inv}, using the Pearson correlation coefficient. The results, shown in figure 3.5, indeed show that the best linearity is achieved with a value of p_{inv} of close to 0.275. Another way of looking at linearity is to plot the extent to which a distance over or underestimates the true distance, taken over the range of true distances. Such a plot is shown in figure 3.6 for a variety of proportions of unvaried sites removed. In general, the flatter the line, the more we may consider the distances additive. By this criteria, removing somewhere in the region of 30% of all sites from the diagonal of **F** results in the highest degree of additivity. At this value of p_{inv}, the true distance is about 3/4 that of the inferred distance, which is still a reasonable estimate given the mismatch of the models. These of how invariant sites models achieve additivity, combined with the importance of additivity, offer one way of looking at why Hadamard conjugation (and we will see later also ML methods) invariant site models tend to give a larger sum of edge lengths than a corresponding gamma distributed model.

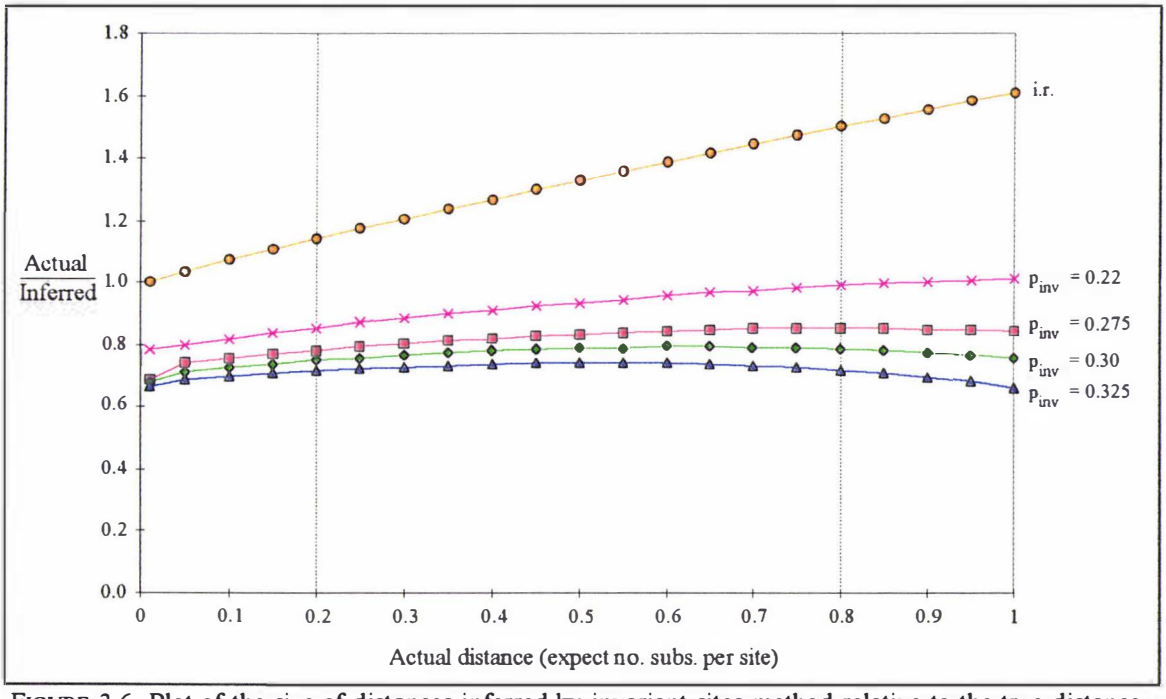


FIGURE 3.6. Plot of the size of distances inferred by invariant sites method relative to the true distance, when the observed sequence data is generated with an exponential distribution of rates across sites (as in fig. 3.4). The closer each curve is to horizontal, the closer the distance is to additive. Most of the distances crucial to estimating the higher order tree structure of the Li and Gouy (1989) 16S-like rRNA data set have a magnitude (measured in no. subs. per site) which falls between the two vertical dotted lines. In this region removing 30% of all sites from the diagonals of the overall **F** matrices, to give $\mathbf{F}_{\text{-pinv}}$, results in very near additive distances.

In conclusion then, removing unvaried sites can dramatically improve the additivity of distance estimates even if the true distribution of rates across sites is quite unlike an invariant sites model (in our example it follows an exponential distribution). As we will see in chapter 5, an invariant sites model can also often fit bimodal data better than a gamma distribution. This suggests that an invariant sites model can offer robustness to the rates across sites problem and we suspect that this degree of robustness may well be in excess of other potential failures of our models other i.i.d. assumptions (for example that all sites have a similar set of relative

substitution rates). We agree with Lake that an unequal distribution of rates across sites is a major concern when analysing ancient functional molecules (but unlike Lake 1994, would not single out rRNA as necessarily being worse than protein coding regions in this regard). With results showing that the invariant sites-LogDet transformation can be nearly additive even though the underlying distribution of rates across sites is an exponential distribution, we feel confident in going towards the next stage of our evaluation of this method.

3.6. IMPORTANT PRELIMINARY STEPS IN ANALYSING SEQUENCES

The data we now analyse are the 28 16S-like rRNA sequences used by Gouy and Li (1989a) to evaluate the statistical support for the archaebacterial tree. Gouy and Li (1989a) ascertained that the alignment of the conserved regions they were using was not a problem, and stated that alternatives to their multiple alignments had little effect upon the tree. We accept their conclusions, but at some future time will be most interested to see if any realignment is suggested by the additional 16S-like rRNA molecules now available.

In this analysis a more stringent alignment than that of Gouy and Li (1989a) is generated by removing all sites where any taxa shows an insertion or deletion (indel), leaving, coincidentally, exactly 800 sites. The edited sequences of Gouy and Li (1989a) are used exclusively throughout the remainder of this chapter, but rarely outside of it. A full list of these sequences is given in the caption of figure 3.12.

The two main motivations for removing all "indels", were (1) the expectation that sites in indel regions would be undergoing substitution more rapidly than sites conserved amongst all taxa, and (2) retaining them implies that the distribution of rates across sites will be different amongst distance comparisons (a violation of the assumption of a stationary distribution of relative rates across sites, see chapter 2). Despite sites in indel regions being useful for resolving finer structure on the tree, their inclusion is expected to result in larger stochastic and systematic errors when estimating the deepest divergences. A third good reason for excluding sites in insertion-deletion regions is that it makes the application of methods of estimating the proportion of invariant sites more straight forward.

The conclusions reached by current methods of analysing rRNA sequences have been the center of recent controversy, with some authors claiming that non-stationary nucleotide compositions may have resulted in serious errors in estimates of the early branchings amongst eukaryotes (e.g. Hasegawa *et al.* 1992 and 1993, Hasegawa and Hashimoto 1993, Hashimoto *et al.* 1994a and 1994b). Further it has been argued by Weisburg *et al.* (1989) and Lockhart *et al.* (1992) that similar problems plague the analysis of both the archaebacteria and eubacteria. Before presenting results of tree building algorithms we believe it is important to gain a "feeling" for the data, especially how it may be violating a tree building methods assumptions. Such information is essential to interpreting the results of any phylogenetic analysis. Firstly we look at ways of assessing the degree of non-stationarity in the substitution process amongst taxa,

something which can be gauged to a large extent by differences of nucleotide composition. Following this we look at methods which can tell us something of the distribution of rates across sites in our sequences, and give estimates of the proportion of sites that may effectively be regarded as invariant. That is, ways of estimating p_{inv}, an essential parameter in the invariant sites-LogDet transformation.

3.6.1 Studying the base composition of rRNA

We begin by comparing the base frequencies in the constant sites versus the parsimony sites (excluding singleton sites). The unvaried sites were [A = 95, C = 42, G = 78, T = 42](totaling 257 / 800 or 32% of all sites), and giving base proportions of [A = 0.370, C = 0.163, G = 0.303, T = 0.163]. At the 551 or 56% of parsimony informative sites the base proportions were [A = 0.184, C = 0.289, G = 0.308, T = 0.219] (measured using the program MacClade 3, Maddison and Maddison 1993). The base composition of the remaining 86 singleton sites is discussed later. The singleton sites are initially ignored as they are a third class which must be nearly invariant, and also a class of which a large percentage could be due to sequencing errors (we check this later). Assuming independence of sites, it is possible to test the null hypothesis that these two sets of frequencies came from the same underlying model. The sampling distribution is a multinomial, and our maximum likelihood estimate of the underlying proportion of "A", for example, is $(f(A)_{unv} + f(A)_{var})/(n_{unv} + n_{var})$ where f(A) is the observed number of A's in the varied and unvaried sites, and *n* is the total number of sites in each set. Doing this for each nucleotide, we can then calculate the expected frequencies for the 8 cells of our contingency table of observed frequencies. Given our assumptions, then under the null model it is expected that the sum of Pearson's X^2 statistic for these 8 cells should be χ^2 distributed with degrees of freedom (d.f.) equal to the number of cells minus the number of constraints minus the number of parameters estimated (there is a constraint on the sum of each set of four cells, and three independent expected frequency parameters are estimated, so test d.f. = 3). The X^2 values for each cell were: Unvaried sites, A: 14.38, C: 6.75, G: 0.01, T: 1.64. Varied sites, A: 8.20, C: 3.85, G: 0.01, T: 0.94. The overall X^2 statistic is 35.767, with associated P value 8.4×10^{-8}. It seems highly unlikely then that the nucleotide composition of the unvaried sites is the same as that of the parsimony informative sites. The result is even more significant than it seems because our test is conservative since it takes just one average value for the 28 sequences considered (a test with near exact probabilities would have to take into account the sampling correlations of the different sequences e.g. due to phylogeny, so that multiple comparisons would not yield inflated overall X^2 values). In comparing the unvaried sites with the varied sites almost all the difference is in the relative frequencies of A and C. Nucleotide A is very rich in the unvaried sites compared to the parsimony informative sites, while C is the exact opposite. This may indicate that the most conserved sites are in the "bulge" and "loop" regions of rRNA as opposed to the more canonically paired (i.e. G:C or A:T pairs) that predominate in the stem regions (e.g. Gutell *et al.* 1985, Vawter and Brown 1993).

The observation that the variable sites have a different base composition to the constant sites, has important implications. LogDet methods will give additive distances amongst taxa

under any i.i.d. and i.r model, but in this case the unvaried sites (which almost certainly contain many "invariant" sites) may be expected to "fool" the LogDet into both misestimating and underestimating the degree of base composition difference between different sequences. It is unclear how robust the LogDet is to this deviation from i.i.d., which may be categorised as different sites having a different set of stationary frequencies.

Further if we expect unvaried sites to contain a significant proportion of invariant or nearly invariant sites, then these sites will essentially be stationary in base composition. Just how important exclusion of unvaried sites from calculations of base composition variability can be is shown in figure 3.7. We measured the difference each sequence is from equal base composition using the Euclidean distance $\left\{\sum_{i=1}^{4}(\pi_i - 0.25)^2\right\}^{0.5}$, where π_i is the proportional frequency of the i-th nucleotide in a sequence. As figure 3.7 shows, when we remove the unvaried sites there emerges a view of large differences in base composition between these species, with the standard deviation of our Euclidean distances nearly doubling. Interestingly the Euclidean distance from equal base composition at the unvaried sites is also quite large (0.179). This means that the base composition of the unvaried sites is nearly as different from equi-frequency, as are the parsimony sites in the thermophiles *Thermus* and *Archeoglobus* (distance approximately 0.21 on figure 3.7). These results also suggest that when we come to remove invariant sites we should be looking a using method (3) or (4) of section 3.5.1.

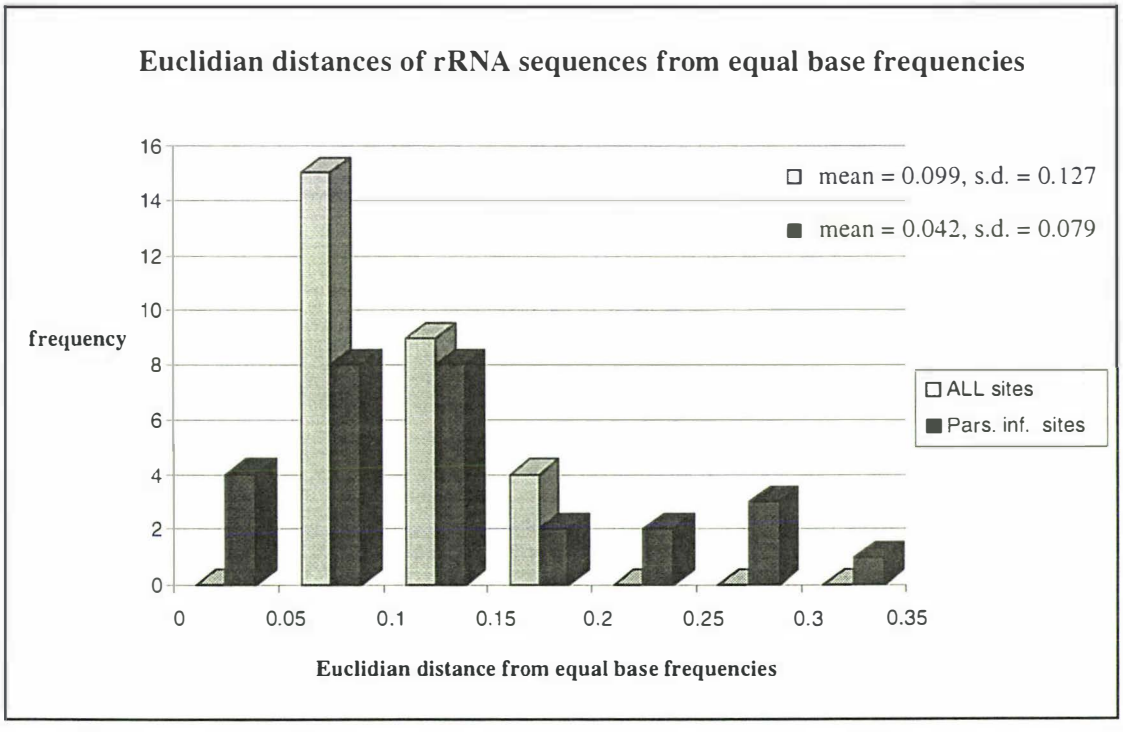


FIGURE 3.7 A histogram of the deviation of the base composition of each rRNA sequence from equal nucleotide frequencies (the data used are the 28 rRNA sequences described in section 3.6) . Note how the dispersion becomes much wider when just the parsimony informative sites (sites known to be variable) are considered. Singleton sites were not included, partly because an unknown proportion of these could be invariant sites which appear variable due to a sequencing error, and because if they are really variable they must be very near invariant in this instance.

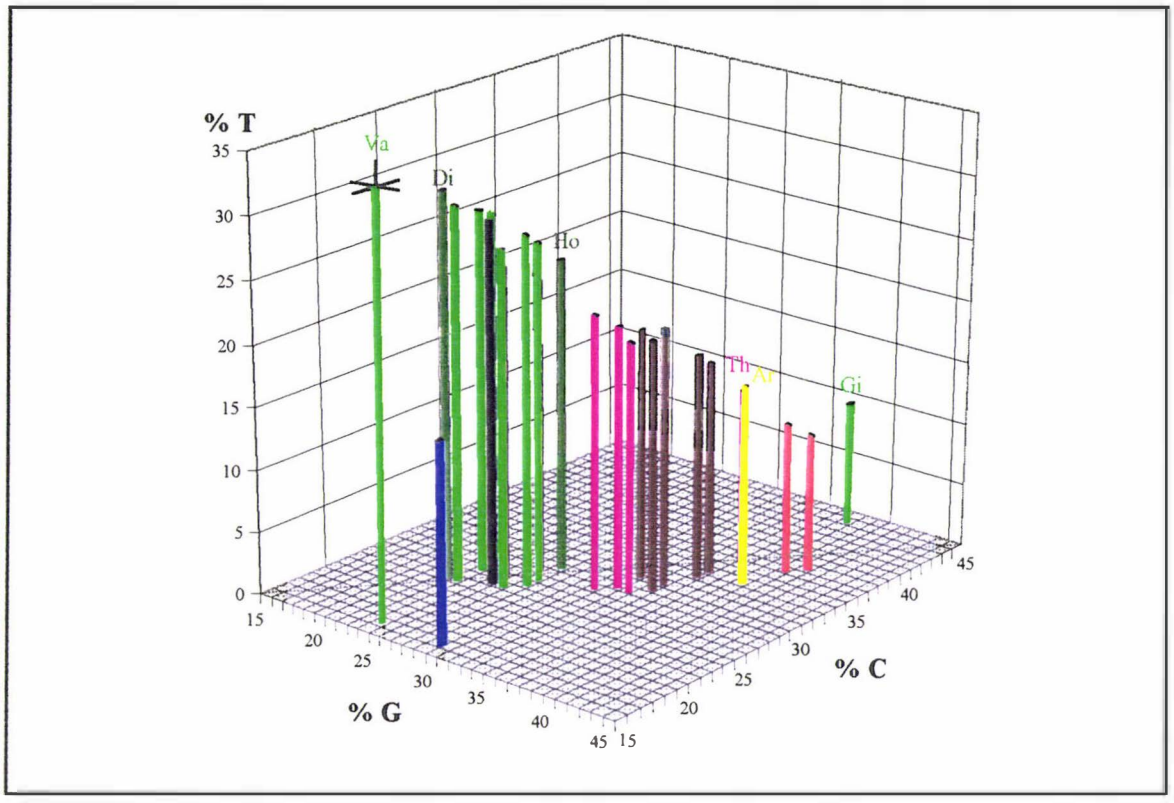


FIGURE 3.8 A three dimensional plot of the nucleotide frequencies of different taxa (consider the tip of each bar to be the nucleotide composition). All information about nucleotide frequencies is contained in the graph (the content of A is equal to 1 - sum of the other three axes) so taxa without identical nucleotide composition cannot have the same bar. The colours represent major groups of taxa (green = eukaryotes), (purple = eubacteria), (red = eocytes), (burgundy = methanogens, expect Archaeoglobus which is yellow). The black column indicates the point of equal base composition, while the blue one is the base composition of the unvaried sites. The labels are: Va = *Vairimorpha*, Di = *Dictyostelium*, Ho = *Homo*, Th = *Thermus*, Ar = *Archeoglobus*, and Gi = *Giardia* (full taxonomic information on these sequences is given in the caption to figure 3.12). If we assume all sites to be independent, then the standard error along any one axis for the percentage of C, G or T is very close to 2% (estimated using binomial distribution)(and fluctuating in the narrow range of 1.7% to 2.3% with the sequences used here). It is therefore easy to imagine the standard error on any one bar tip to be approximately a sphere with radius 2%, as shown by the axes at the base composition point for *Vairimorpha*. (the standard error for the unvaried sites is slightly less than 3% at its greatest breadth, due to fewer sites being used). Clearly most points are significantly differentiated from one another.

Figure 3.8 shows another useful way of presenting the base composition of sequences. The top of each bar indicates the base composition as a point in three dimensional space, with the bars serving principally as a means of giving perspective to these points. An interesting property of this graph is that if GC and AT content vary concomitantly (i.e. frequency of G = C and A = T, e.g. f(g) = f(c) and f(a) = f(t)), then such sequences would fall perfectly in a line from the top left corner to the point 5% above the lower right corner in figure 3.8. Clearly most of the base compositional change amongst these sequences follows this trend, with the exception of two dramatic outliners, these being *Giardia* and *Vairimorpha*. This trend may be largely due to the predominantly canonical pairing in the stem regions of these molecules (e.g. GC pairs with triple hydrogen bonds, and AT pairs with double hydrogen bonds), and higher GC is structurally correlated with greater thermal stability (see Gutell *et al.* 1985).

The "deviant" sequences not following a trend of f(g) = f(c) and f(a) = f(t) are far apart, and show opposite deviations from the general GC trend. *Vairimorpha* favours G over C, while *Giardia* favours C over G. This mismatching of GC contents suggests that the base compositions of these species cannot be readily identified with increasing or decreasing thermal stability. Indeed neither of these species is expected to be under much selective constraint due to thermal stability of rRNA, as both are mesophilic (like their relatives) living at temperatures below 45 degrees Celsius. It is interesting however that the departure of these species from the equal GC line, is in the direction which preserves nearly equal purine / pyrimidine (i.e. AG vs CT) content. This observation is compatible with a modification to the DNA repair enzymes such that they still have the ability to discriminate between nucleotide chemical classes, but not so well within classes. Apart from these outliers, note the generally tight clustering of the species groups on this figure. For these sequences there is little deviation within the more recent eukaryotes, or within the eocytes, methanogens and eubacteria (except for the thermophilies *Archeoglobus* and *Thermus* whose GC contents again appear subject to selection for higher GC to confer thermal stability), but often distinct differences between these major groups. Only a part of these differences can be explained in terms of the temperature these organisms live at.

The results in figure 3.8, combined with the results of section 3.4, suggest that we should expect that nucleotide composition differences will cause underestimation of the distance between *Giardia* and the thermophiles which are prominent amongst the prokaryotes, while *Vairimorpha* will generally be repelled from the prokaryotes (and *Giardia*) due to its base composition (when using standard distance transformations). The archaebacteria may show some clustering together due to their generally similar base composition, hence potentially hiding evidence for the "eocyte" tree. The methanogen *Archeoglobus* is experiencing some "attraction" to the eocytes, while base composition is causing *Thermus* is be "attracted" more towards archaebacteria than to any other eubacteria. The eukaryotes (*Giardia* and *Vairimorpha* aside) are the only species to have near equal nucleotide frequencies (i.e. near equifrequency, as indicated by the black bar in figure 3.8). This type of plot offers a valuable alternative to building a tree based on base composition distances (e.g. Lockhart *et al.* 1994), as it retains all information about the composition distance between different species. It does not lose important information such as the fact that *Giardia* and *Vairimorpha* are deviating substantially from the general f(g) = f(c) and f(a) = f(t) trend. Further, it is easy to estimate if two species have significantly different base compositions, by considering the ellipsoid of errors about each point. For example figure 3.8 shows the axes of such an ellipsoid about the base composition of Vairimorpha, which is clearly significantly different to all other species.

The blue bar in figure 3.8 indicates the average base composition of the unvaried sites in these sequences. It is clearly unlike the base composition of the parsimony sites of any other sequence, suggesting that these sites have quite different evolutionary forces acting upon them. Note that this situation is probably more extreme than that seen in mtDNA, where the base composition of the constant sites more closely reflected that of the of the variable sites (see section 1.9.2),.

Perhaps even more startling than the average base frequencies at the parsimony informative sites, were the parsimony singleton sites of *Vairimorpha* and *Giardia* (by singleton site we mean sites where all taxa have the same base except the taxon indicated). *Giardia* had unique substitutions [A = 1, C = 11, G = 3, T = 1], while *Vairimorpha* had changes [A = 9, C = 1, G = 2, T = 10], suggesting an even more biased substitution process than that indicated by averages across variable sites. The extreme base composition of these substitutions, concordant their greatest frequency in these two long branched taxa, suggests few of them are due to sequencing errors. Thus these do appear to be genuinely extremely conservative site substitutions. Accordingly it is expected that a strong substitution bias has operated in these two lineages over a considerable period of time. This in turn suggests that relatives of these species will have also experienced the same strong underlying substitution bias over a long period of time, which should still be evident in their base composition. The only other species to show more than two singleton substitutions where *Crithidia* [2, 2, 1, 2], *Physarum* [1, 0, 0, 5] and *Thermus* [1, 1, 1, 1] (in order of A, C, G, T). *Crithidia* and *Thermus* show singletons consistent with long periods of independent evolution, but no strong bias. However *Physarum* looks unusual and despite the sample being small and the possibility of sequencing errors, its substitution biases will be interesting to check in future with other genes. These findings highlight further difficulties to be expected in resolving deep parts of the eukaryotic tree. Interestingly the archaebacteria, and the eocytes in particular, show very few singleton changes. Assuming these species to be anciently diverged, this reinforces the view that they are in generally evolving at a slow rate, something evident also when trees are built from all sites (e.g. Olsen 1987, Lake 1988).

3.6.2 Five different types of method to infer the number of invariant sites

As already seen these rRNA sequences are expected to have large site to site differences in evolutionary rate, and for this reason Lake (1994) has recently suggested that LogDet methods should not be used on such data. In section 3.5 we devised a way of overcoming this effect, and showed it is expected to result in a marked improvement. Due to the quite distinct base composition of the unvaried sites, we should be using either method (3) or (4) from section 3.5.1 to give robustness to unequal rates across sites. We now address the question of estimating the precise number of constant sites to remove in order to yield the best additivity of transformed distances. Illustrated next are four different types of method, including a new application of capture-recapture techniques applicable to rRNA. By using four methods with different assumptions, we hope to obtain by conciliance, a robust estimate of the number of sites which should be removed before making distance corrections or constructing ML trees.

3.6.3 A new capture-recapture method suitable for rRNA

The first method we consider is a new application of a capture-recapture model to sequence data, and was inspired by the work of Seber 1982, and Sidow *et al.* 1992. Firstly we separate our sample of sequences into two disjoint subsets. By exclusive we mean that assuming we knew the unrooted evolutionary tree, then all paths between sequences in set 1, share no edge in common with paths between sequences within set 2. This separation is important because we can treat

evolutionary change in set 1 as independent of that in set 2 (and it implies monophyly, via a unique common ancestor for at least one of the two sets of sequences). We then estimate the quantities n_1 (number of sites that show any change in set 1), n_2 (number of sites that show any change in set 2) and m_2 (the number of sites which show change within both sets) (see table 3.3). Treating changes at a site in an analogous way to tags on animals, then our estimate of the number of the number of variable sites is,

$$\hat{N} = n_1 n_2 / m_2 \qquad (3.6.3\text{-}1)$$

a statistic sometimes called the "Petersen estimate" (Seber 1982, p.59). The sample variance of $\hat{N}$ (via the delta method approximation) is, $V[\hat{N}] = n_1 n_2 (n_1 - m_2)(n_2 - m_2) / m_2^3$. So our estimated faction of invariant sites is $\hat{p}_{inv} = (c - \hat{N}) / c$, where c is the total number of sequence sites (or codons if amino acid sequences are used). Consequently the standard error of $\hat{p}_{inv}$ is,

$$\text{s.e. } \hat{p}_{inv} = \frac{1}{c}\sqrt{\frac{n_1 n_2 (n_1 - m_2)(n_2 - m_2)}{m_2^3}} \qquad (3.6.3\text{-}2).$$

The decision of how to separate the data into two sets is governed by two main factors: (1) Being confident at least one of the groups is monophyletic). (2) Keeping the probability of observing changes about equal in the two sets (an important point to help minimise the variance of when we estimate p_{inv}). For our rRNA sequences separating eukaryotes from prokaryotes achieved this aim very well. Table 3.3 shows the resulting numbers of changes observed within each set and the proportion shared. Our estimate of p_{inv} is thus 0.269 (with s.e. 0.013), so this method (with 95% confidence) estimates that 24.3 - 29.5% of sites are invariant (with 95% confidence). As Seber (1982) and Sidow *et al.* (1992) point out, this estimate is the maximum likelihood estimate if we make no further assumptions about the model (in this case of evolution).

Table 3.3 Table of changes in prokaryotes, eukaryotes and both groups

	Euk.			
Prok.	change	no ch.		
change	**284**	97	381	n_2
no ch.	152	?	?	?
	436	?	$\hat{N}$?	?
	n_1			

The four central cells in this table will be independent if sites are evolving i.i.d.. Therefore the ratio of entry m_2 (in bold) divided by column total n_1 must equal $n_2 / \hat{N}$. Rearranging $m_2 / n_1 = n_2 / \hat{N}$ we have $\hat{N} = n_1 n_2 / m_2$, which here gives $436 \times 381 / 284 = 584.92$. Thus $\hat{p}_{inv} = (800 - 584.92) / 800 = 0.2692$.

It is worth examining the assumptions of this capture-recapture statistic in light of the biology of molecular sequences. We now rewrite the assumptions required for $\hat{N}$ to be a reasonable estimate of N (see Seber 1982, p59), in terms of sequence evolution:

(1) The sequence is of constant length and made up of only correctly aligned homologous sites

(2) All sites have the same probability of substituting in the first sample (excepting those that are strictly invariant).

(3) A site changing in the first set does not affect the probability of a site changing in the second set

(4) The second sample of site substitutions is a simple random sample, i.e. each of the (N choose n_2) possible samples is equally likely.

(5) Sites do not loose their "marks" in the time between two samples (see later).

(6) All marks are reported on recovery in the second sample (see later).

Assumption (1) is clear. Assumption (2) is clearly unlikely to be correct in a coding region, because of unequal rates across sites (i.e. different intrinsic rates of substitution). Assumption (3) is just as potentially important. If a covarion model is operating, then generally we expect that some of the sites changing in one part of the tree, will become invariant in another part (Fitch 1971). Consequently sites showing change in both parts of a tree will be rarer than expected if the covarion model was not operating. There is a high likelihood that a covarion model is playing a significant role in the evolution of these molecules (given the observed slow substitution rate of many sites, the clear hierarchical structure of changes, and the vast period of time, see discussion section of chapter 2) and hence we expect m_2 to be biased downwards due to this factor (i.e. an underestimate), making N an overestimate, and in turn biasing the estimate of p_{inv} downwards. Unequal rates across sites also violates assumption (4). That is with highly catchable sites, we underestimate the number of variable sites, and so overestimate p_{inv}. However in the last section of this paragraph we explain why this overestimation is usually a desirable feature. Lastly non-independence between sites will also result in a violation of assumption (4), which we expect will principally affect the reliability of our estimate of the s.e. of p_{inv}, almost certainly making it an underestimate. Assumptions (5) would seem to hold as the mark in this case is the nucleotide state, which could only be lost if we were including "indel" regions, which we are not. Assumption (6) could be violated by sequencing, or other errors, in defining the final data. The expectation is that these types of errors, if random, will tend to bias the estimate of p_{inv} downwards.

Given that the estimate of p_{inv} by this method is in good agreement with estimates by four other methods (as we will soon see), it seems likely that our two probable causes of bias (unequal relative rates at sites, and a covarion effect) are to a large extent canceling each other out. This need not be the case with other sequences, where one or the other effect may dominate, depending at least partly on the degree of divergence amongst the sequences. One possible way to reduce the error of unequal substitution rates would be to eliminate sites showing many changes on a reliable estimate of the tree (and hence unlikely to have any of their number showing no changes), and base the estimate of unvaried but variable sites on the remaining set of more slowly evolving sites. The properties and biases of this modification need further study.

An extension of this method would be to classify sites into a number of rate classes, and make a capture-recapture estimate of p_{inv} from each, then sum up all these estimates to get the overall estimate. Other possibilities are given an estimate of the distribution of rates across sites from some other method (say Γ distributed with shape parameter 0.8), predict how many unvaried sites there should be between two groups (this can easily be done if we approximate the continuous distribution with a few discrete rate classes, using the mean rate of each to represent that class). There are even some analytic results relevant to this issue (see for example capture-recapture estimates which assume that "catchability" follows a Pearson type 3 distribution on p. 177 of Seber 1982). In any case, as far as is apparent here, the standard estimator works reasonably well as an overall estimate of how many sites should be treated as invariant (given that we do not also distinguish between the substitution rates of the variable sites).

We now look more closely at sample size bias in estimating N and the variance of our estimate of N. Under the hypergeometric geometric distribution $\hat{N}$ is a maximum likelihood estimate of N (Seber 1982, see also Sidow *et al.* 1992). However it is also a biased estimator. An approximately unbiased estimator (exactly unbiased when $n_1 + n_2 \geq N$) is,

$$N^* = \frac{(n_1 + 1)(n_2 + 1)}{(m_2 + 1)} - 1 \qquad (3.6.3\text{-}3)$$

while an approximately unbiased (exactly unbiased when $n_1 + n_2 \geq N$) of its variance is,

$$V[N^*] = \frac{(n_1 + 1)(n_2 + 1)(n_1 - m_2)(n_1 - m_2)}{(m_2 + 1)^2 (m_2 + 2)} \qquad (3.6.3\text{-}4)$$

(Seber 1982). Consequently our unbiased estimate, p^*_{inv} becomes (1 - N*/c), while the standard error of p^*_{inv} becomes $1/c\sqrt{V[N^*]}$. Applying these formulae to our rRNA data gives a very slight increase in p^*_{inv}, and decrease in $V[p^*_{inv}]$ (both in the fourth significant place) relative to the ML estimators (for this data $n_1 + n_2 \geq N$, clearly holds). Clearly in this example the validity of our assumptions is our main concern regarding the accuracy of our estimates.

The capture-recapture method of Sidow *et al.* (1992) has recently been used to infer the proportion of invariant sites in some anciently diverged proteins involved in chlorophyll and bactriochlorophyll synthesis (see Lockhart *et al.* 1995). The values of n_1, n_2 and m_2 obtained were 38, 20, and 15 respectively, in a comparison amongst a set of proteins with similar function (the protochorophyllide reductase subunits, chL + bchL). However values of 64, 43, and 32 (respectively) where obtained when some more diverged proteins were added to the previous data set (chL + bchL + the chlorin reductase subunit, bchX) (in both cases the entire sequence length was 121 codons). It is important to know if the estimate of *N*, the variable sites, goes up significantly when bchX genes are added; if so it is a strong indication that these sequences have a different set of sites able to vary than those in chL and bchL. Because our numbers of captures and recaptures are much smaller than our previous rRNA example, it is necessary to look closer at the test statistic. For this data $\hat{N}$ (chL + bchL) = 50.67, with s.e. 5.09, while N*(chL + bchL)

= 50.19, with s.e. 4.65, while for the expanded data set we have, $\hat{N}$ (chL + bchL + bchX) = 86, with s.e. 10.73, while N*(chL + bchL + bchX) = 85.67 with s.e. 10.29. There is little difference when using the unbiased estimators, although the difference in the means appears slightly larger and more significant (also since it seems nearly certain that $n_1 + n_2 > N$, then the N* estimator should be unbiased). Seber (1982, p62) mentions that a confidence interval about 1/N* is usually more symmetric than that about N*. So for the (ChL + bChL) data 1/N* = 0.099, with s.e. 0.0018, so finding ± 2 s.e. of 1/N* and then inverting we have, a 95% C.I. on N*(ChL + bChL) as 42.3 - 61.6 (vs. 40.9 - 59.5 for the direct approach). For N*(ChL + bChL + bchX) a 95% C.I. by the inverse method is 69.1 - 112.8 (vs 65.1 - 106.3), so making some difference (this method applied to the rRNA data analysed previously subtracts just 0.1% from the previously given limits of the C.I.). Since these two confidence intervals are non-overlapping it seems likely the means really are different. Unfortunately our assumption of normality may not be justified, Seber (1982) suggests that if p or (1-p) (where $p = m_2 / n_2$) is less than 0.3, then n_2 should be at least 80 for the normality assumption to hold. Otherwise if $p > 0.1$ a binomial approximation can be used, or if $p < 0.1$ use a Poisson approximation.

In this example a direct test of the difference of means is in order, so x - y = (85.67 - 50.19) = 35.48, with variance = Var[X] + Var[Y] - 2Cov[XY]. Estimating the covariance requires counts of how many sites estimating n_1, n_2 and m_2 are shared in common by the two sets of sequences. This test becomes complicated to do exactly and approximations could result in inaccurate estimates of the true sampling distribution of the mean difference in the number of covarions estimated, unless n_1, n_2, m_2, (n_1-m_2) and (n_2-m_2) are all large (one would suspect at least 50 each), so that the distribution is particularly close to multivariate normal. Given the ease with which simulations can be performed with modern computers (or even a spread sheet such as Excel), we do not follow this avenue further here, but recommend the use of simulations.

As we will see later, the capture-recapture estimate is apparently giving reasonable estimates of the number of variable sites for the rRNA data. For the purpose of making LogDet distances more additive, it would seem that the standard Peterson estimate applied to all sites, is giving an estimate of p_{inv} which is biased upwards enough to compensate most fully for unequal rates across sites. To help explain this, note that under a Γ distributed model there are no truly invariant sites, and with many recaptures this could be established. Yet this would then give an estimate of p_{inv} that was quite useless (i.e. $p_{inv} = 0$) for improving the additivity of LogDet or other i.r. distances. This same overestimation of p_{inv} is almost certainly also happening with all the other methods when rates vary across sites, but the model is assuming just i.r. variant and invariant sites. In these instances the bias towards overestimating p_{inv} is apparently benevolent to our purpose of improving distance additivity. This type of overestimation of shape parameters when assuming an incorrect form of distribution, probably occurs with all rates across sites models. In these too it appears to be a benevolent bias, so that the end result appears to be distances that are much closer to additive.

3.6.4 Using "observed" numbers of changes to infer p_{inv}

This second method is based on the expectation that the number of changes per site is Poisson distributed (e.g. Fitch and Markowitz 1971). Using parsimony, we can estimate the number of substitutions that have occurred at each site. Although this will tend to give underestimates, with the rRNA data we suspect it to be introducing little error. This is because estimates of the true tree show it to be quite evenly branched, with most sites showing far fewer than the maximum possible number of changes that could be detected by parsimony (see Wakeley 1993 for an assessment of the bias in these estimates). Using the optimal LogDet tree (see figure 3.12) which is our best estimate of the true phylogeny, in place of the maximum parsimony tree, should help give a more reliable estimate of the number of substitutions per site. We present these numbers as the black columns in figure 3.9. To this data various distributions and mixtures of distributions are now fitted. The essential assumption is that given an intrinsic rate λ, the expected number of substitutions per site with this rate follows a Poisson distribution with mean λ (e.g. see Fitch and Markowitz 1970, Uzzel and Corbin 1971, Wakeley 1993).

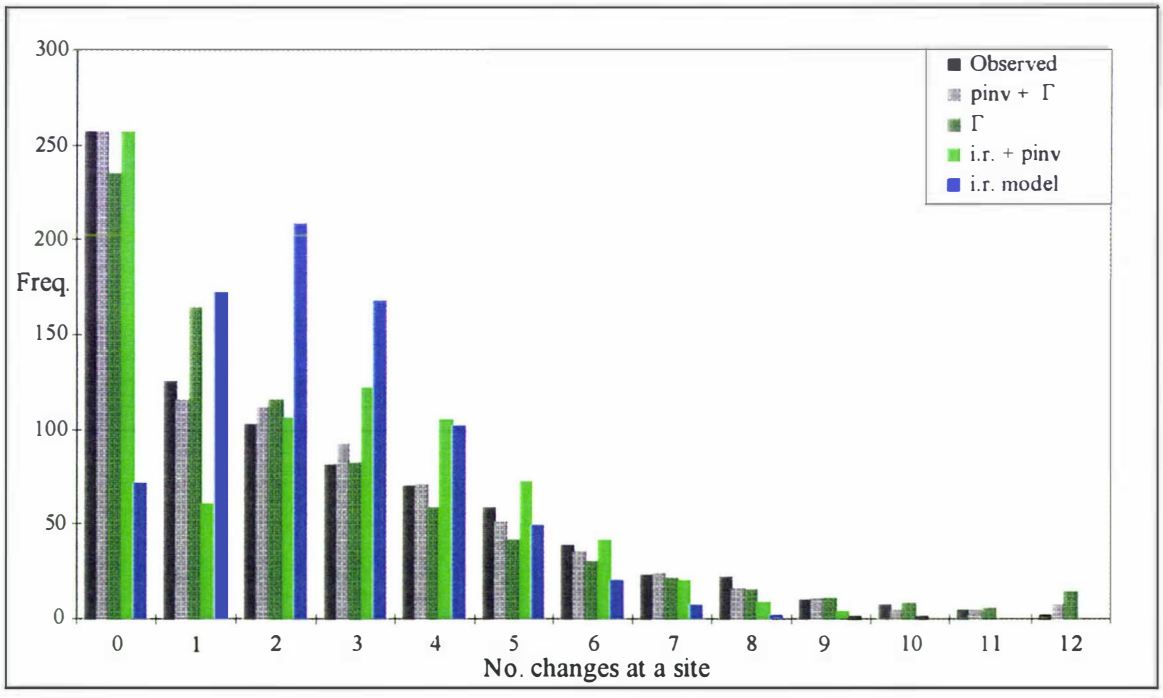


FIGURE 3.9. A plot of the observed number of changes per site for the 16S-like rRNA data of Gouy and Li 1989a, and in different colours the expectations of different models. The invariant sites plus a gamma distribution of variable sites model fits very well, while the remaining models show various inadequacies (e.g. the gamma model shows poor fit of the sites with one change, the invariant sites model shows a bimodality not seen in the data, while the i.r. model is poorly fitting every where). The observed numbers of changes per site were measured by the parsimony criterion upon the optimal tree shown latter in figure 3.12. The expected number under different models were predicted by the probability distributions described in the text, with all free parameters optimised so as to minimise the G^2 ln likelihood ratio fit of observed to expected data.

Letting the number of substitutions per site be x, and letting the intrinsic rate (mean expected number of changes for a set of sites with identical rates) be λ, then $P[X = x] = \lambda^x e^{-\lambda}/x!$ under a standard Poisson model (Stuart and Ord 1987 p. 160, see Fitch and Markowitz 1970 for an

application). Given P[X = x] then the expected number of sites showing x changes is just P[X = x] $\times$ c (the sequence length). Having both the observed and expected numbers of sites showing x changes we can measure the agreement of these two sets of numbers using either the Pearson X^2 statistic, or G^2 (the log likelihood ratio statistic, explained further in chapters 5 and 6). The next step is to choose λ to minimise either of these statistics (we will use G^2) and this can be done using an iterative procedure. The resulting expected number of changes per site under the Poisson model are shown in figure 3.9 as the blue columns, which clearly do not fit so well.

Table 3.4 Fit of different distributions to parsimony inferred number of changes per site

Model	d.f. (no.p)	G^2	$\chi^2_{0.01}$	ln L.R.	p_{inv}	k_1 (k_2)	$\Theta\Gamma$	λ_1 (λ_2)	p_1
Observed	9 (0)	-	-	-					
Γ + Γ	4 (5)	2.87	13.28	-		0.619 (29.9)	2.58 (0.15)		0.69
Γ + i.r.	5 (4)	3.17	15.09	0.30		4.61	0.79	0.249	0.64
i.r. + i.r.	6 (3)	24.33*	16.81	21.46*				0.533(4.43)	0.51
Γ + p_{inv}	6 (3)	7.89	16.81	5.02	0.216	2.45	1.26		0.78
Γ	7 (2)	36.24*	18.48	28.35*		0.97	2.52		
i.r. + p_{inv}	7 (2)	135.69*	18.48	127.80*	0.299			3.46	0.70
i.r.	8 (1)	665.55*	20.09	657.66*				2.42	

The fit of various models to the data shown in figure 3.9. The degrees of freedom (d.f.) shown are after grouping the 3 highest observed rate classes in order to give a better approximation of fit statistics to asymptotic χ^2 expectations (no. p, is the number of parameters optimised in each model). The G^2 goodness of fit statistic indicates that of the models shown, only the gamma (Γ) mixed with another distribution fits within expected sampling error (models rejected by this statistic at the 99% level are marked *, while the next column gives χ^2 critical values). All the models shown here are submodels of a mixture of two gamma distributions. Likelihood ratio tests (ln L.R.) are thus valid between the most general model and each submodel. The likelihood ratio test fails to reject either the Γ + i.r. model or the Γ + p_{inv} models as being worse fitting at the 99% level (rejection again indicated by *). The drop in the G^2 statistic when the Γ + p_{inv} model is generalised to the Γ + i.r. model suggests a significant increase in goodness of fit, however the Γ + i.r. it does not have the explanatory power of the Γ + p_{inv} model; its better fit may just reflect an inadequacy in the assumed form of the distribution of the sites that can change (i.e. the Γ distribution assumption). The choice of model on such fine points might also be influenced by our expectation that a few of the singleton sites are sequencing errors. Another, nonnested model, made up of p_{inv} + i.r.(λ_2) + i.r.(λ_2) also fits very well and is discussed in the text. The optimised model parameters are the proportion of invariant sites, p_{inv}, the parameters of each Γ distribution, k and $\Theta\Gamma$, the substitution rate in each i.r. distribution, λ_i and p_1 the proportion of sites belonging to the first component of the two part distributions.

If we assume λ_i (here the absolute rate at the *i*-th site) to be randomly drawn from a gamma distribution (with shape parameter *k*, and scale parameter P), then it is possible to integrate the Poisson predicted number of changes for each infinitely narrow rate class, to give the expected number of substitutions per site. This yields the often used negative binomial distribution, so

$$P[X = x] = \left(\frac{k+x-1}{k-1}\right)\left(\frac{P}{P+1}\right)^x\left(\frac{1}{1+P}\right)^k, \qquad (3.6.3\text{-}2)$$

where x = 0, 1, 2 ..., the number of observed events, *k* is the shape parameter of the underlying gamma distribution (k > 0), and P = ΘT (P > 0) is a scale parameter from the underlying gamma distribution (the mean number of substitutions per site per site is given as kΘT) (Stuart and Ord 1987 p169, see Uzzel and Corbin 1971, and Wakeley 1993 for applications). When *k* is not an integer, then the first term in this formula can be evaluated as Γ(k + x) / (Γ(k)!x), where Γ(z) is the gamma function (a generalisation of the factorial function) applied to z. It is straightforward to add a proportion of invariant sites (p_{inv}) to either of these distributions as all such sites belong to the 0 changes class. It is also easy to mix and match models by adding up their proportional expectations, while optimisation of the component distributions free parameters was accomplished with quasi-Newton and checked with conjugate gradient optimisation methods (e.g. see Minoux 1986).

The results of fitting a number of models to the observed number of substitutions per site are given in table 3.4. The invariant sites / i.r. (i.r. + p_{inv}) model gives the estimated number of invariant sites as between 26% and 30% depending on the measure of fit used, i.e. X^2 or G^2, and whether the 3 highest rate classes are grouped. Work by Wakeley (1993) suggests that this estimator of p_{inv} will usually be biased slightly downwards, but still reasonably accurate, as long as there is an adequate sampling of diverse sequences. The optimal shape parameter when fitting a gamma distribution of rates across sites model was 0.97 (similar to the figures obtained by Hadamard conjugation and ML analyses of the same type of data in chapters 2 and 5). However, the only models fitting within expected sampling error, when assuming all sites to be independent, mixed the gamma with a second distribution. The sum of two weighted gamma distributions model fitted best of those evaluated (table 3.4), however the gamma plus an i.r. rate model fitted almost as well with 1 fewer parameter, while the gamma plus invariant sites model was also a very good fit (an Akaike type model selection would favour the Γ + i.r. model, although the improvement in fit is just significant at the 5% level over the Γ + p_{inv} model, e.g. see Miller 1990 for criteria in picking a model). We favour the last of these three models because it fits prior expectations better than the other two somewhat contrived models (i.e. what biological model predicts a mix of two gamma distributions?). Lastly the sum of two Poisson distributions (i.r. + i.r.) plus a proportion of invariant sites model (not included in table 3.4) was also found to fit very well, and has only 4 parameters (G^2 = 4.88, optimised parameters are p_{inv} = 0.23, λ_1 = 1.53, λ_2 = 5.12, proportion of λ_1 = 0.42). This model and its subtypes were considered by Fitch and Markowitz (1970) and Uzzel and Corbin (1971). The very good fit of this last model is

relevant to the penultimate section of this chapter where another approach to making LogDet type transformations robust to unequal rates across sites is described.

When using the X^2 goodness of fit statistic to select a distribution of rates across sites much more weight was given to the fit of the rarer cells showing high rates of change, relative to using the G^2 statistic. Consequently the fit of the i.r. model was over twice as bad as given in table 3.4, while the mean number of changes inferred for variable sites rose for the two worst fitting models (which have trouble explaining the prevalence of rapidly evolving sites). An 95% confidence interval on the number of invariant sites under the p_{inv} + i.r. model was 0.265-0.335 when the G^2 criteria of fit was used, or 0.225-0.295 by the X^2 criterion (the confidence intervals estimated assuming a binomial marginal distribution to the inferred number of invariant sites were nearly identical). So given our understanding of the invariant sites-LogDet transform, this data suggests that we remove approximately 25 to 30% of unvaried sites. Even the favoured model (Γ + p_{inv}), infers about 22% of all sites invariant, and this is expected to be an underestimate (e.g. see figure 3.6) if the aim is to make the LogDet optimally additive without separating the variable sites into rate classes.

Lastly note that this method can be modified to remove the effect of parsimony tending to underestimate the number of changes at a site, by using probability calculations to directly infer the expected parsimony length of sites under a certain model. To do this we would start with a reliable estimate of the weighted tree, predict **s**(T) for a specific mechanism of substitution, and sum up the probabilities of all sites having the same parsimony length on the model tree (for some mechanisms we could use a Hadamard conjugation, else the more usually likelihood calculations to estimate the probability of all site patterns). Multiplying this probability by the sequence length would then give the expected number of sites with that parsimony length. These statistics of observed to expected parsimony length might also be useful in diagnosing the fit of ML models, by checking the ability to explain the frequency of sequence sites of a certain length (see chapter 6).

3.6.5 Inferring p_{inv} with a ML model of sequence evolution

The third method of estimating the proportion of invariant sites is finding the value which maximises of the likelihood of a set of sequences as evaluated by some model. For this purpose we used the stationary 5 parameter i.i.d. and i.r. model of Felsenstein, as implemented in the program DNAML 3.5 (Felsenstein 1993). Others that have used this sort of method include Reeves (1992), Churchill *et al.* (1992), and Sidow *et al.* (1992). The calculations evaluate the likelihood of the data as a mix of variable and invariant sites, analogous to the probabilities for **s**(T) under the invariant sites models in chapter 2. We seek the maximum likelihood tree, and the proportion of sites treated as invariant which will maximise its likelihood. The intricacies of these calculations were touched upon in chapter 2, and are given again later in this section and in chapter 5. For this section we use an option in DNAML 3.5 to make these calculations (the option is to allow different rate classes, specify two rate classes, set the first to any positive value and the other to zero along with the proportion of sites to assume invariant). This

program, as compiled, had a limit of 19 sequences, so this many widely separated sequences were selected from the 28 shown in figure 3.12. The results of this analysis are presented in figure 3.10; like earlier estimates, this method too suggests that about 30% of all sites should be regarded as invariant under an p_{inv} + i.r. model.

It is difficult to know if this estimate is biased, with respect to the optimal additivity that could be achieved for LogDet distances. Results elsewhere in this thesis (chapter 5) suggest that when the model is in some way moderately inadequate, and cannot explain an excess of certain site patterns, then the proportion of invariant sites estimated often increases as a compensatory factor. This possibility seems reasonable since our results clearly show that a stationary i.i.d model cannot explain the very unequal base compositions between sequences. However, we also observe situations where if the model is strongly inadequate in some crucial way, then the above trend need not hold and can even be markedly reversed, with the result being a much lower estimate the number of invariant sites (see chapter 5). It would be interesting to re-evaluate this data with an ML model that allowed for non-stationary base compositions (especially a 12 parameter per edge model which is analogous to the LogDet).

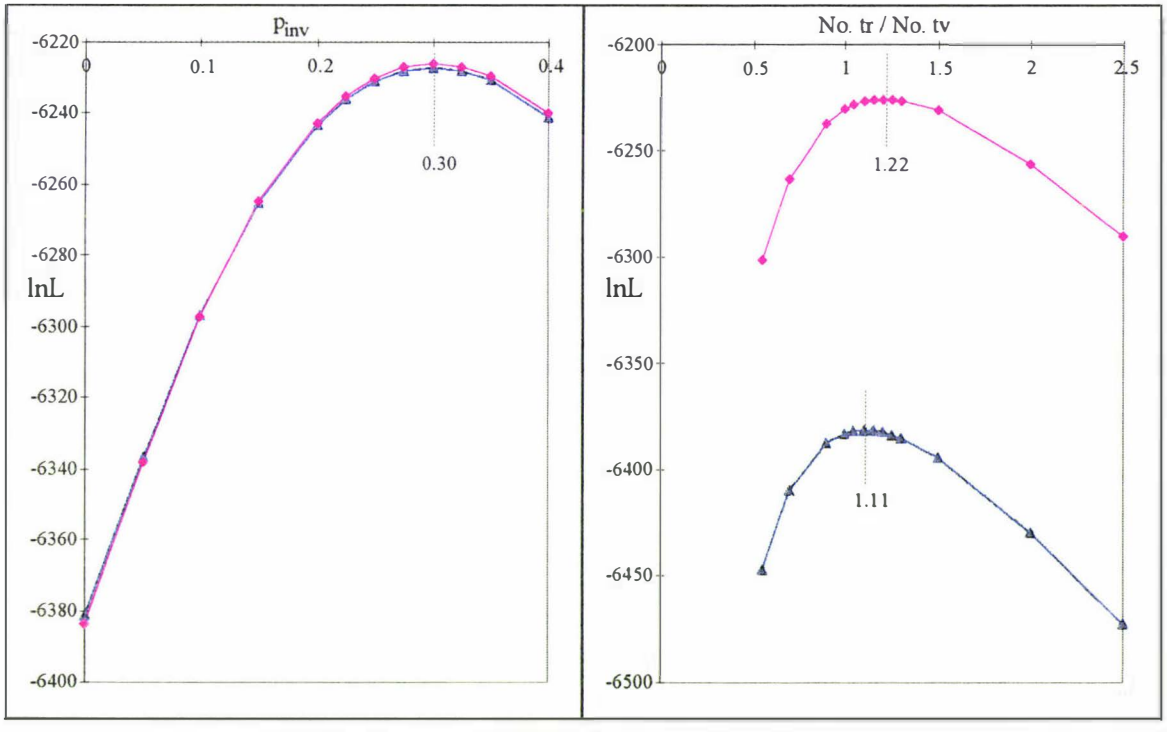


FIGURE 3.10 (a) A plot of log likelihood vs proportion of sites modeled as invariant. The line with triangles represents the curve obtained with a tr / tv parameter of 1.11, while the line with diamonds represents that obtained with tr / tv set to 1.22. (b) Curves showing how the optimal tr / tv ratio increases as we take account of invariant sites (the upper curve marked by diamonds, and with higher likelihood, is when 30% of sites are treated as invariant, the lower curve is with no invariant sites). In this example the tr / tv ratio is having a little effect upon the estimated proportion of invariant sites.

In the case of the DNAML model it would also be desirable to take into account the distinct base frequencies of the unvaried sites, which are probably largely composed of invariant sites. The likelihood of the constant sites, both unvaried and invariant sites mixed together, under such a model is just $\{p_{inv}\pi_j + (1-p_{inv})(s(T)_j)\}$, while the likelihood of the varied sites is $(1-p_{inv})s(T)_i$

(where π_j is the proportion of invariant sites with base frequency *j*, $s(T)_j$ is the probability of a variable but unvaried site with base *j*, and $s(T)_i$ is the probability of the *i*-th pattern under some i.i.d. model of evolution without invariant sites). If such a model where implemented in place of the current model in DNAML we would expect the likelihood of the data to increase by at least tens if not hundreds of lnL units. We base this expectation on the X^2 lack of fit between the base frequency of the varied and the unvaried sites (calculated earlier in section 3.6.1). Our anticipation is that this quantity will be a lower bound on the improvement. The X^2 test should gauge the increase in likelihood at the unvaried sites quite well, but the increase for the varied sites will probably be much larger, since these likelihoods involve many multiplications of edge transition matrices, which are conditional upon base frequencies.

Models taking into account distinct base compositions of the invariant sites will have increased robustness. As an example, consider what happens when base compositions in the variable and the invariant sites are stationary, but skewed in opposite directions, e.g. π_{var} = [0.1, 0.4, 0.4, 0.1], while π_{inv} = [0.4, 0.1, 0.1, 0.4]. Overall if 50% of sites are invariant, we will observe near equal base composition in such sequences. Applying DNAML to these sequences, it would effectively revert to using the Kimura 2ST model, when it should be ignoring the invariant sites base composition and using a 5 parameter model for the remaining variable sites.

3.6.6 Estimating p_{inv} by directly measuring additivity of distances on a tree

Given that we are wanting to modify distances to make them more additive, it makes sense to find the proportion of sites assumed invariant which does this directly. Probably the most exact way is to used a method such as generalised least squares (GLS, see chapter 5) which measures the significance of errors on distances, taking account of their correlated errors, to give a statistic which is asymptotically ($c \rightarrow \infty$) distributed as a χ^2 variable (see chapter 5). This requires a variance-covariance matrix of distances; a delta method approximation to the covariances of the LogDet distances is currently being sought . If we ignore the correlations between distances, then we have weighted least squares which we have already implemented for LogDet on another data set (see figure 3.3). There, the inferred number of invariant sites (about 0.15) differs from that obtained with all other methods probably due to correlations between distances not being taken into account and because only four, perhaps atypical taxa where used from a different alignment. GLS estimation of p_{inv} should be achievable for up to about thirty taxa (the major step being inverting a 30 × 30 element covariance matrix), although the evaluations would probably be limited to a few trees selected as near optimal by other methods.

Other distance methods such as ordinary least squares or minimum evolution do not appear to be very promising for this type of fitting. This is because they are not associated with a fit statistic that takes account of increasing variances and covariances as unvaried sites are removed (more on this in chapter 5). Hasegawa *et al.* (1985) probably used a fitting procedure of this type; they are very frugal in their description and it is not clear exactly what they did.

3.6.7 The Bealey theorem inequality

While there is no simple way to predict exactly the expected frequency distribution of parsimony length under a particular model, there has recently been developed, an inequality which goes some way towards meeting this aim. This is the Bealey theorem of Steel *et al.* (1993a). It states that under the 2-state Poisson model, the probability of a site showing *m* or more substitutions (denoted P[m ≥ x]) on the true tree, must be less than or equal to a simple function. Specifically,

$$P[m \geq x] \leq P[1] \sum_{j \geq x} \frac{v^{j-1}}{j!}, \text{ (where is v equal to P[1]/P[0]}^2\text{)} \qquad (3.6.6\text{-}1)$$

This inequality can be used to find an approximate lower bound on how many sites can be regarded as invariant. One way to do this is by requiring all the observed site lengths strictly meet this inequality (although we must ignore the inequality if the observed number of sites with *x* changes is zero). It is best to measure the parsimony length of sites on the best estimate of the true tree, which is not necessarily the shortest parsimony tree. If all site lengths do not initially meet the constraints of the Bealey theorem, then we vary the quantity P[0] (proportion of sites showing no change) until the constraint is meet. Table 3.5 gives an example of these calculations, using some four taxon transversion data analysed more extensively in chapter 5. With few taxa this lower bound on p_{inv} tends to be very loose (compare the lower bounds of table 3.5, which are 0.7 and 6.7%, with about 15% when the same data is analysed with ML in chapter 5) and also quite dependent on which tree is selected as "true".

Table 3.5 Predicting minimum number of invariant sites with the Bealey inequality

	Observed data		Bealey theorem predictions		
			no inv.	75 inv.	
x	Obs[m = x]	Obs[m ≥ x]	Inf.[m ≥ x]	Inf.[m ≥ x]	
0	787	1352	1261.4	1352.6	(1277.6 + 75)
1	457	565	803.3	880.6	
2	108	108	339.3	416.6	
3	0	0	104.3	145.4	

Here *x* is the number of changes per site, while Obs [m = x] is observed number of sites showing *x* substitutions as measured by parsimony upon the "eocyte" tree for the transversion data used in figure 5.2. We have given the predictions of the Bealey theorem as inferred number (P[m ≥ x] × sequence length) of sites that should show *x* or more changes. By the inequality of equation 3.6.6-1 the inferred number should be less than the observed number (not the case with the underlined number), and it was necessary to infer that 91 sites were invariant before this inequality could be meet by all site patterns. This gives a lower bound on the number of invariant sites of 91/1352, i.e. at least 5.5% of sites are invariant. Using the same methods and data but measuring Obs[m = x] on the "archaebacteria" tree gave a prediction of more than 0.7% invariant, while the halobacterial tree inferred at least 6.7% invariant.

Table 3.6 Using the Bealey theorem to infer p_{inv} for the 28 16-S like rRNA sequences

	Measured on tree from LogDet NJ					Measured on Max. Pars. tree		
Inv.			0	202				191
x	Obs[m = x]	Obs[m ≥ x]	Inf.[m ≥ x]	Inf.[m ≥ x]	x	Obs[m = x]	Obs[m ≥ x]	Inf.[m ≥ x]
0	399	800	458.28	1026.89	0	399	800	926.04
1	165	401	259.28	759.99	1	166	401	664.00
2	88	235	93.28	594.99	2	100	236	498.00
3	75	135	24.04	385.24	3	70	148	304.06
4	40	65	4.79	207.48	4	38	73	153.00
5	19	27	0.78	94.49	5	20	33	64.75
6	8	7	0.11	37.04	6	2	14	23.51
7	3	5	0.01	12.69	7	3	6	7.45
8	2	2	0.00	3.85	8	2	3	2.09
9	1	0	0.00	1.04	9	0	1	0.52
10	0	0	0.00	0.25	10	0	0	0.12

As with the previous table x is the number of changes reconstructed by parsimony on a tree. Obs[m = x] indicates the number of sites with *x* changes ([m ≥ x] number of sites with at least *x* changes), while inferred (Inf.) is the Bealey theorem inequality value. The first set of site lengths was measured on the LogDet tree from figure 3.12, which we consider the most reliable estimate of the true tree. The row marked Inv. indicates how many invariant sites are included when calculating the inequality which is first meet by all sites with 202 invariant sites (25.25%). To the right are estimates made with tree lengths measured on the maximum parsimony tree for the unweighted observed R /Y coding of the data (191 or 23.88% invariant sites for the inequality to hold).

Applying the Bealey theorem inequality to the transversional changes in the Gouy and Li (1989a) data set, we estimate that at least 202 sites (25.3%) are invariant, when taking the tree of figure 3.12 (estimated using the LogDet transform) to be the true tree. Alternatively if we take the maximum parsimony tree for this data (shown in figure 2.14) as the true tree, we infer that at least 191 sites (23.9%) are invariant.

The numbers returned by this approach can be termed an approximate lower bound for two reasons. Firstly because there are no expected values calculated, we cannot use a measure such as X^2 to define when the randomly fluctuating observed proportions best meet the theorem's predictions. Secondly as we have already seen, a distribution of rates across sites with a distinct tail (e.g. the inverse Gaussian or the gamma distribution) tends to produce a much higher probability of sites with many changes than under the corresponding invariant sites plus identical rate model. With the 28 taxon rRNA data, it was just these sites which were the last to meet the inequality (and they meet the inequality in reverse order of their parsimony length). Unlike a model which is fitted by the G^2 statistic, for example, the absolute requirement of meeting an inequality gives these sites absolute leverage in determining what proportion of invariant sites are estimated. A situation where the Bealey theorem method may overestimate the proportion of invariant sites would be a data set with more taxa and / or a longer tail of rapidly evolving sites.

For this reason we consider the Bealey theorem method interesting, but generally not as robust or reliable as the other methods used in this section. The close agreement of the Bealey theorem and the other methods is encouraging, but may be largely due to coincidence in this instance (notice for example how large, and hence loose, the Bealey theorem predictions in table 3.6 become as the number of changes at a site decreases). Consequently the Bealey method should be used with caution, and preferably never be the only method used to infer a proportion of invariant sites.

3.6.8 Summary of diagnosing this 16S-like rRNA

It is clear from these studies that we must treat a proportion of sites as being invariant and that the base composition of these invariant sites is quite distinct from the very nonstationary base compositions of the variable sites. As such it seems most appropriate to use an invariant sites-LogDet transform upon this data prior to tree estimation. The estimates of the proportion of invariant sites were predominately in the range of 0.25 to 0.3, and as such were quite consistent given the different assumptions of the four distinct estimators applied to all 28 sequences. In this section and elsewhere in the thesis, these data when fitted to a Γ model predict the shape parameter k to be approximately 0.8 to 1. Even if the Γ where a better model, then figures 3.5 and 3.6 suggest that 25 to 30% of unvaried sites is an appropriate proportion to remove in order to improve the additivity of LogDet distances.

3.7 FIELD TRIALS OF THE INVARIANT SITES-LOGDET TRANSFORMATION

The purpose of this section is to evaluate the utility of the invariant sites-LogDet upon a set of data which has clearly evolved by non-stationary substitution process with a marked inequality of rates across sites. Like any good field trial a method needs to be evaluated side by side with a strong competitor. Our analyses here make a paired comparison between the invariant sites-LogDet method and the invariant sites-Jukes-Cantor distance correction, so we can identify where and to some extent why different transformations are leading to different conclusions about evolutionary history. The Jukes-Cantor correction is a very popular measure with this type of data (e.g. Bruns *et al.* 1992, Hinkle and Sogin 1993). It is often argued to be an appropriate transformation since the transition / transversion ratio in this type of data is low and because this measure has a lower sampling variance than any other commonly used distance correction.

To avoid setting the Jukes-Cantor distance up as a "straw man" many other parallel assessments were made using other distance transformations in the programs DNADIST, form Phylip 3.5, and PAUP* (Swofford in press). These distances include the Kimura 2P method with a gamma distribution of rates across sites (Golding 1983, Jin and Nei 1990), and the general time reversible distance with a Γ distribution of rates across sites (this thesis section 3.3.2, and Waddell and Steel in preparation). Also tried were a variety of ML methods (including removal of constant sites) and also maximum parsimony methods. Some of the results are surprising and

tend to indicate that the invariant sites-Jukes-Cantor transformation followed by the neighbor joining tree selection algorithm was possibly one of the better performers (more on this later). Where we noticed a substantial difference of any other method from the results of the Jukes-Cantor / neighbor joining combination it is noted in the text.

An important question is how an invariant sites-LogDet analysis might change our perceptions of the "tree of life" which has for the past decade been inferred primarily with methods based upon simple i.r. and stationary i.i.d. models of evolution. One possibility is that the invariant sites-LogDet tree will be distinctly different to the trees previously obtained. An alternative possibility is that large parts of the tree will be shown to have very low bootstrap support, thus bringing into doubt the resolving power of the data. This second conclusion was often reached by Peter Lockhart (pers comm.) in analyses of datasets such as those in Lockhart *et al.* (1994) (although these were often run after removing all constant sites, or all parsimony uninformative sites). Poor resolving power was also the expectation of statisticians such as Terry Speed (pers comm.), who suggested the method would have critically high sampling variance due to its generality. If it turns out that the LogDet transform shows similar resolving power to the Jukes-Cantor and other simple i.i.d. distances with these highly diverged and relatively short rRNA sequences, it has quite acceptable sampling errors. In order to make the most of this opportunity to look at major aspects of organismic evolution in a new light, we have chosen to test six specific prespecified hypotheses.

In this following section we use the term "methanogen" to include the halobacteria, and any archaebacteria showing evidence of having at least had ancestors with the enzymes required to produce methane. This grouping is well supported by biochemical features and previous sequence analyses of 16S-like and 23S-like rRNA (e.g. see Olsen and Woese 1989). The term "crown group" relates to plants, animals, fungi, red and brown algae, plus their immediate protist relatives (e.g. Sogin 1991). The term "middle eukaryotes" is used for the mitochondrial protists that do not appear to be closely related to the crown group, which in this study are taken to comprise *Dictyostelium*, *Physarum*, and *Crithidia* (see figure 3.12 for full details of the other taxa in this study).

3.7.1 Six prespecified hypotheses about the "tree of life"

We evaluate the following hypotheses which were selected prior to data analysis; indeed all of these hypotheses are discussed in others research, and all are presently controversial. Our hypotheses to "test" are:

(I) Are the archaebacteria a single cluster on the unrooted tree? This grouping was suggested by Woese and Fox (1977), with additional support claimed in many recent publications (although not proving the strict monophyly, i.e. exclusive ancestry of this group e.g. Hennig 1966, this thesis conveniently labels this hypothesis "archaebacteria monophyletic"). The most prominent alternative hypothesis claims good evidence that thermophilic sulfur metabolising bacteria (so called "eocytes") are sister taxa to the eukaryotes (Lake 1986, Wolters and Erdmann

1986, Lake 1987, Rivera and Lake 1992)(this implies eukaryotes and eocytes will form one half of a partition on an unrooted tree).

(II) Are the parasitic Microsporidia the most anciently diverged eukaryotes? This is suggested by their 5.8S-rRNA being part of the 23S-like rRNA molecule (Vossbrinck and Woese 1986). Alternatively does this distinction go to the diplomonad lineage (often characterized by *Giardia*) as suggested by the rRNA analysis of Sogin *et al.* (1989)? Leipe *et al.* (1993) identified this as a situation where compositional bias may be misleading standard methods, yet nearly all published trees present *Giardia* as deepest branching of the eukaryotes (ironically many of these still originating from Dr Sogin, e.g. Brul and Stumm 1994, fig. 3). Hasegawa *et al.* (1992, 1993) and Hasegawa and Hashimoto (1993) have also voiced their concern that base composition is misleading rRNA trees of early eukaryotic evolution.

(III) Can we reject the traditional view of the strict monophyly of the slime molds *Dictyostelium* and *Physarum*, as claimed in Sogin (1991)? If so, how does Sogin's alternative hypothesis of *Physarum* being more distantly related to the "crown group" than *Dictyostelium* fare.

(IV) There is the old question of whether animals are sister taxa to plants or to fungi, or are these latter two closest relatives? The grouping of plants and fungi is the traditional view, popular before the last decade of sequence analysis.

(V) The grouping of the methanococcal and methanobacterial archaebacteria appears in trees estimated from 23S-like rRNA (my own unpublished analyses of the 23S sequences aligned by Gouy and Li 1989a, Burggraf *et al.* 1991, although these authors made no comment upon this grouping). This grouping has also appeared in RNA polymerase trees (Garret *et al.* 1994, Klenk and Zillig 1994). The question: Is this grouping consistent with the 16S-like rRNA data?

(VI) Is the thermophilic genus *Thermus* (which is allied with the non-oxygenic photosynthetic green non-sulphur bacteria e.g. *Chloroflexus*, in most 16S-like rRNA trees) more deeply diverging than the oxygenic photosynthetic cyanobacteria? Fossil stromatolite bacterial forms, photosynthetic pathways, and more recently trees from DNA dependent RNA polymerase genes have been used to argue for the more ancient divergence of cyanobacteria (Klenk *et al.* 1994). Conversely trees built from 16S-like rRNA have tended to put many thermophiles, including *Thermus*, nearer to the root of the eubacteria than cyanobacteria. Doubts currently loom over the validity of the rRNA trees due to a "GC" bias drawing *Thermus* towards the archaeobacteria (e.g. Lockhart *et al.* 1992).

Hypotheses (I), (II) and (VI) are considered good candidates for a marked difference between the invariant sites-LogDet transform and previous analyses, since the sequences emanating from the internal edges of the phylogenetic tree defining these hypotheses show distinctly unequal nucleotide content (as shown in figure 3.10). It is important to note that while we call these hypotheses, because of their contentious nature each can comprise up to three

mutually exclusive hypotheses. In these cases it is necessary to be aware of alterations in significance levels, a question we address later.

3.7.2 Using the bootstrap as a guide to statistical support.

To help evaluate the support for our hypotheses I-VI we use a variant of bootstrapping (Felsenstein 1985, Penny and Hendy 1985), which places emphasis on a local part of the tree. If a tree has a stable structure, then many points of uncertainty will be limited to local rearrangement about internal edges in the optimal tree (nearest neighbor interchanges in the language of tree searching, e.g. Swofford and Olsen 1990, Swofford 1993). Most of our hypotheses I-VI probably fall into this category. By plotting the sum of the three partitions in the bootstrap replicate trees that correspond to the three rearrangements possible about a specific edge in a tree, we can see how the support for these hypotheses varies as different distance transformations are applied to the data. Further, by taking the sum of the three partitions about an internal edge in the bootstrapped trees, we can construct a test of the hypothesis that "the edge in the true tree is either this edge, or an edge only one nearest neighbor interchange different." That is we will accept that a hypothesis about the relative relationships of taxa can be framed in terms a nearest neighbor interchange, if the taxa (or groups of taxa) of interest are separated by no more than one internal edge in more than say 90% of all bootstrapped trees. This type of test is an novel example in the spirit of Sanderson (1989), who argued there is much benefit to be gained by examining which loose groupings where very frequent in bootstrapped trees.

The results of a bootstrap analysis with particular emphasis on our six hypotheses are shown in graphical form in figure 3.11. Note that for each hypothesis we have two tests, the second one being that if the hypothesis does not gain overwhelming support in its own right, then what is the likelihood that the true relationships are one of three possible nearest neighbor interchanges about a single internal edge in the optimal tree. Here we are "testing" six sets of hypotheses, and later in chapter 6 we consider the multiple test problem in more detail.

To accommodate variation of rates across sites we took out a specified proportion of constant (unvaried) sites (p_{inv}). We did this such that the number of constant columns with nucleotide *i* which were removed was $p_{inv}\pi_{unv}$, where π_{unv} is the frequency of the *i*-th nucleotide amongst just the unvaried nucleotide sites (overall 32.5% of the sites were unvaried). Taking out constant sites in proportion to π_{unv} is an important factor since the base frequencies of the unvaried sites are significantly different from those at which changes occur (section 3.6.1). Following this we applied a distance correction then used the neighbor joining algorithm to a build a tree (via the program "Neighbor" in Phylip 3.5, Felsenstein 1993). Our selection of this distance algorithm was based upon its good statistical efficiency in estimating trees in simulations of 20 to 30 taxa (using distances with similar sampling error to those used here, Charleston 1994, chapter 6). Its selection was also based upon its computational efficiency (Studier and Keppler 1988), which allowed us to perform 1000 bootstrap replicates upon each distance matrix. Preliminary investigations by Charleston (1994, chapter 5) suggest that

neighbor joining can handle violations of the model as well as many other commonly used methods (maximum likelihood was not included in these comparisons). A check of the reliability of neighbor joining for assessing bootstrap support is made with the related minimum evolution criterion, and the agreement was high for this data set (full details in section 3.7.4).

As a further check we analysed the distance matrices with an ordinary least squares criteria (Cavalli-Sforza and Edwards 1967, as implement by Felsenstein 1993 in the program "Fitch") and a weighted least squares criteria (Fitch and Margoliash 1967, also implemented in Fitch). The support as gauged by bootstrapping with these two tree selection procedures was practically identical to that obtained by neighbor joining when the bootstrap support for an edge exceeded 70% (the agreement was less when negative edges were allowed in the optimal trees, see Kuhner and Felsenstein 1994 for general discussion on this issue). For all other partitions the smaller number of bootstrap replicates performed with these methods (100 samples) generally agreed to within 5% to 10% of the values given by neighbor joining.

The bootstrap has been somewhat controversial throughout its usage in phylogenetics (e.g. see Sanderson 1989, Hillis and Bull 1993). There is a general feeling from simulations that bootstrap results can be regarded as conservative estimates of support in real situations, because they tend to be so in simulations. However the bootstrap assumes independence between sites, and yet in rRNA sequences there are correlations (often strong between base pairing sites) which will tend to exaggerate statistical support measured under the assumption of independence (Felsenstein 1985). We will discuss these and other issues further in chapter 6, but for the present it is worth being aware of this possibility.

3.7.3 Support for our six hypotheses with the invariant sites-LogDet transform

The bootstrap support for the prespecified hypotheses are shown in figure 3.11. The region where analyses are usually performed is on the extreme left of each subplot, i.e. allowing no accommodation for unequal rates at different sites. Our previous assessment of the distribution of rates across sites in this data set suggests that at least 20% of unvaried sites should be removed, with good evidence (sections 3.6.2 - 3.6.6) that the optimal proportion to remove is 25 to 30%. This view is reinforced by the earlier numerical finding that if the true distribution of rates across sites is Γ then the shape parameter should be approximately 0.9 to 1 (see figure 3.9, and table 3.4). Even if this were the distribution, then the results in figures 3.4-3.6 suggest the removal of 25% to 30% of unvaried sites to optimise the additivity of the invariant sites-LogDet transform. (All the more suggestive since the favoured $\Gamma + p_{inv}$ model of table 3.4 would suggest at least this many sites be treated as invariant to optimise additivity).

For each hypothesis our interpretations after this analysis are:

(I) The results indicate that taking into account both GC bias and more particularly variable rates across sites drops the support for the archaebacterial tree. With 0% of constant sites removed our bootstrap analysis reaches a similar conclusion to the simple edge length confidence interval analysis of Gouy and Li (1989a), suggesting significant support for the

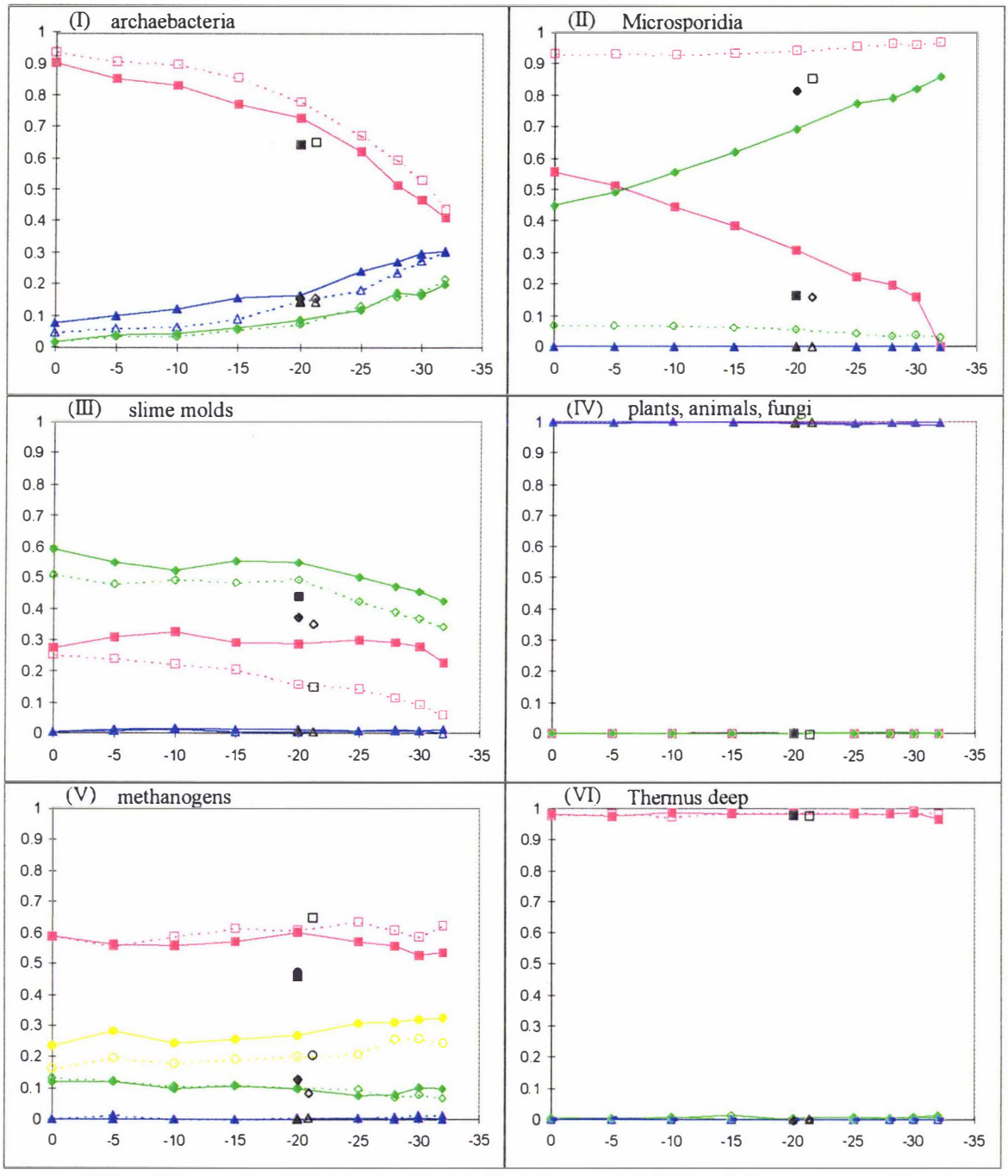


FIGURE 3.11. Relative support for the six a priori hypotheses presented in the text (labeled as I-VI). The horizontal axis shows the proportion of constant sites removed p_{inv} (removing more sites increases the adjustment for unequal rates of change at different sites). The y-axes denote the frequency with which a the bootstrapped trees supported a specific hypothesis. The solid lines and symbols indicate trees reconstructed after the LogDet transformation has been applied, dotted lines the equivalent support with the standard Jukes-Cantor correction. In general the red squares mark what is assessed to be the most "popular" hypothesis (prior to this analysis, and based upon recent publications), while green diamonds mark a prominent alternative hypothesis. (I) squares (red), archaebacterial tree, diamonds (green) eocyte tree, triangles (blue), halobacterial tree, (II) squares, *Giardia* deepest of all eukaryotes, diamonds Microsporidia deepest, triangles *Giardia* and Microsporidia together and deepest, (III) squares, *Physarum* deeper than *Dictyostelium*, triangles the reverse, diamonds *Physarum* and *Dictyostelium* together, (IV) squares, fungi and animals together, diamonds plants with animals, triangles plants with fungi, (V) squares methanococcals deeper than methanobacterials, triangles the reverse, diamonds methanococcal-

methanobacterial grouping, yellow circles *Archeoglobus* deepest of methanogenic archaebacteria considered, (VI) squares, *Thermus* deepest, diamonds cyanobacteria deepest, triangles *Thermus* and cyanobacteria together. The solid black symbols at -20 denote support for the various hypotheses after removing 20% of all sites (only constant columns removed), plus removal of the 5.6% of the sites showing most change on the best supported tree, followed by the LogDet correction. Hollow symbols denote support when just 20% sites are removed, and the 5 parameter ML distance transformation in Phylip 3.5 is used.

archaebacterial tree at the 95% level. However we cannot accept this analysis in light of the evidence for a highly skewed distribution of rates across sites. When we evaluate the data with 20 to 30% of unvaried sites removed, we must conclude that the archaebacterial tree has lost much of its support. The total frequency of the three alternatives archaebacterial, eocyte plus halobacterial tree was over 95% even with 28% of sites treated as invariant, so it seems likely that the correct arrangement is one of the three shown. This result is in contrast with Lake (1988) whose analyses suggested a much more fragmented archaebacterial group.

(II) The compensation for unequal nucleotide composition, and for unequal rates across sites has a major effect upon the support for Microsporidia being the most anciently branching lineage amongst the eukaryotes (i.e. the sister taxa to all other living eukaryotes). Indeed use of the invariant sites-LogDet turns the result around almost as far as is possible (from *Giardia* first having high bootstrap support to Microsporidia first having strong bootstrap support). Interestingly, with no constant sites removed the tree placed the microsporidium sequence deepest, whereas on bootstrapping the same data *Giardia* was slightly (but more significantly than can be explained by error in estimating the bootstrap proportions) favoured as earlier in the consensus tree of bootstrap replicates. A possible cause of this is that the largest distances in the tree are from the prokaryotes to *Vairimorpha* (the microsporidian) and this combined with their dissimilar base compositions lead to the largest biases in the whole tree. This could then result in the microsporidium being pushed further away from the prokaryotes than *Giardia* and hence away from the root. In contrast *Giardia* is involved in the second largest set of distances in this data, and its base composition sees an attraction towards the prokaryotes in general, but just as importantly a repulsion from all the other eukaryotes. Given these two factors acting in concert it is not so surprising that distance trees to date, have tended to place *Giardia* deepest, and that this deep placement is even more forceful as distance estimates become larger when taking into account variation of rates across sites.

The hypothesis of Microsporidia and Diplomonadea (here represented by *Giardia*) being specifically related (albeit very anciently) receives no support (not one tree in all the 18,000 bootstrap replications). In the case of the Jukes-Cantor correction this could well be due to a strong bias of different base compositions repel (section 3.4.2), while in the case of the LogDet it could be this and / or partially a sample size bias. This example shows nucleotide compositional bias worse than that reported in Weisburg *et al.* (1989), since unlike their analysis, concentration on the conserved sites alone still sees the Jukes-Cantor distance selecting what we consider is most likely an incorrect edge in this part of the tree. With both distance

corrections, the two alternative hypotheses of "Microsporidia first" or "*Giardia* first" account for 99% of all replicate trees, with 28% of invariant sites removed.

(III) In the case of the "slime molds", the hypothesis of *Physarum* and *Dictyostelium* together has most support, but the support for *Physarum* being a more ancient divergence is not far behind (especially with the LogDet correction). The hypothesis of *Dictyostelium* being more distantly related to the "crown" group than *Physarum* seems highly unlikely by either method. These three hypotheses together do not account for more than 78% of all trees (with p_{inv} = 28%) with the LogDet transform, and even less with the Jukes-Cantor correction (just 50%). Consequently we expect that the resolution of the branching order of the "middle eukaryotes" will be quite challenging. Although we did not set out to test this hypothesis, in figure 2.12 we see reasonable support placing *Dictyostelium* outside the grouping of plants, animals and fungi. This is interesting in that one of the first papers to suggest base composition bias problems was Loomis and Smith (1990) who made this suggestion because they felt that their protein sequences placed *Dictyostelium* in the right place (which from their analysis was next to plants, see also Hasegawa *et al.* 1992 who reach the same conclusion). Based on both the invariant sites LogDet analyses, very similar placement for the ML, parsimony and other distance analyses (e.g. figures 3.13, 3.14) and the fairly standard base composition shown in figure 3.8, it seems likely that if there is a misplacement of *Dictyostelium* in the rRNA analyses, it is not due simply to overall base composition. (Note though, we cannot exclude the possibility of biased substitution causing problems, without examining more closely the actual substitution their location on the sequence, and the rate classes they belong to).

(IV) Surprisingly this data combined with these methods gives strong and unwavering support for plant and fungi being closest relatives. This result contradicts a previously published analysis (Wainright *et al.* 1992) based on 16S-like rRNA, which claimed significant support for the animal-fungi association (although many similar studies have been equivocal). Some phylogenetic analyses of slowly evolving protein sequences claim support for the animal-fungi grouping (e.g. Baldauf and Palmer 1993), while others (Gouy and Li 1989b, Sidow and Thomas 1994) claim significant support for the grouping of animals and plants together. (However both these later studies are disturbing in that they used outgroup sequences from prokaryotes or very early gene duplications to root the tree). I have noticed however that at least one tree for heatshock protein 70 gene sequences (HSP70, a molecular chaperone located on membranes) shows a grouping of plants and animals to the exclusion of fungal sequences (e.g. see Boorstein *et al.* 1994, although the original authors did not comment upon this feature of their tree). What is particularly persuasive is that all three subfamilies of HSP70, which are apparently found in all eukaryotes with mitochondria, repeat this same grouping. These results are from a gene which is arguably the most alignable and conserved of all known ubiquitous protein sequences (although caution is called for since a trypanosome HSP70 sequence groups with fungi in one of the protein subfamilies).

This strong support for plants and fungi together immediately raises the question of whether large subunit (23S-like) rRNA sequences, which almost certainly had the same evolutionary

history, support this result. Our own reanalysis of Li and Gouy's (1989) conserved aligned regions of large subunit rRNA's gave 95% bootstrap support to the plant-fungi grouping, when the invariant sites-LogDet transform was applied (20% of constant sites removed)(very similar results were obtained with other methods). Clearly much remains to be learned about the process of molecular evolution, and it will be particularly interesting to see why these analyses of conservative sites contrast so strongly with inference from the larger data set used in Wainright *et al.* (1992). (Note however that Rodrigo *et al.* 1994 criticise the Wainright study, suggesting the claims in this paper are not well supported). We will examine this same hypothesis of plant, animal and fungi relationships again later in this chapter using other methods of analysis.

(V) Moderate support for the methanococcal lineage earlier than the methanobacterials comes from most transformations of this 16S-like data. The grouping of methanococcals and methanobacterials has low support, steady at about the 10% level. Thus it is still reasonable to consider that this data set and analysis do not completely refute the grouping of methanobacterials and methanococcals evident in other analyses, including those of 23S-like rRNA sequences. In addition, either of these groupings plus the alternative that methanobacterials branch immediately prior to the methanococcals, occurred in only 65% of all bootstrap trees, suggesting that still other rearrangements are quite possible. These three hypotheses are not separated by just one edge difference in the optimal tree (rather the taxon *Archaeoglobus* comes between these taxa) so this last cited support has a slightly different meaning to rearrangements about a single internal edge. (Rather it is rearrangements of these taxa over two internal edges, or alternatively rearrangements about an internal edge when the position of the taxon *Archaeoglobus* is not important).

Thus it seems likely that sorting out the exact branching order of the methanobacteria will also be challenging. In this regard we take the opportunity to make a second (not a priori) test of a hypothesis that Woese *et al.* (1991) argued for, namely that *Archaeoglobus* is quite deeply nested in the methanobacterials, being a sister taxa to the methanomicrobials and the halobacteria. Opposite to this conclusion were many earlier analyses that suggested *Archaeoglobus* was near the root of the methanobacterial taxa as then known. (the yellow line shows support for this hypothesis) Interestingly the LogDet transform did not drop the support for *Archaeoglobus* deep, one which Woese *et al.* (1991) argue is an artifact of base composition bias. Their own evidence for placing *A. fulgidus* as sister taxa to, or even within the methanomicrobial-halobacteria group is based upon a procedure called "signature analysis." This is essentially a process of looking for specific motifs (sometimes single base substitutions) which characterise most members of a group. In this instance it will be interesting to explore which hypothesis (if either) is correct, as an analysis of the A'-A'' subunits of RNA polymerase (Garret *et al.* 1994) also suggests that *Archaeoglobus* is not a sister taxa to just the methanomicrobials and halobacteria (when their tree is pruned of additional taxa not in our analysis).

(VI). The final hypothesis is answered in the affirmative, *Thermus* indeed appears to diverge deeper than the Cyanobacteria. The similarity of support by both the Jukes-Cantor model and the

LogDet model suggests that in this instance unequal base composition was not a major factor (the four singleton changes in the *Thermus* sequence support this conclusions as they are to each different base).

Before concluding this section, we consider how the programming error (see end of section 3.3.1) in calculating the LogDet in the programs "Trees" and "Prepare" (Penny *et al.* 1993) may be distorting bootstrap results reported here. We have checked all the results presented here with early versions of the program, PAUP* (Swofford 1995) known to be calculating LogDet distances accurately. Figure 3.12 compares support by the two programs in one instance, with 20% of constant sites removed. Overall the support was very similar. Because these two programs are using different pseudo-random numbers, the expected difference for bootstrap support (b.s.) is the difference of two binomial distributions with the same expected value, and 1000 trials each. Accordingly the expected bootstrap disagreement in figure 3.12 is 2.2% when b.s is 50%, dropping to 2.0% when b.s support is 70 or 30%, then dropping below 1.3% when b.s. support is 90 or 10%. Clearly some of the differences in support are significant. Certain trends were detected with larger bootstrap sample sizes, and these are now described in terms of how they alter the support for sets of hypotheses (I)-(IV). The flawed LogDet distance was underestimating the distances between sequences, and was generally giving similar results to the true LogDet distance but with p_{inv} approximately only 3/4 as large. The bootstrap support difference with these two distances was often only marked with 30% or more of all sites treated as invariant. With all unvaried sites treated as invariant the previously well supported edge grouping eukaryotes, and also the edge grouping eubacteria suddenly dropped from 100% bootstrap support to the low 80's. Accordingly we suspect that it was at this point that sampling variances became distinctly larger and may underestimate the true support. With more than 30% of sites treated as invariant there was also a sharp increase in the number of negative determinants occurring in the bootstrap samples.

Support for the archaebacteria monophyletic fell off slightly more quickly with the correct LogDet transforms than is shown in figure 2.11, so with 32.5% (or all) unvaried sites treated as invariant the support for the archaebacteria tree was nearly random at just 35%. In addition the support for the eocyte tree was nearly always as high as that for the halobacterial tree.

In figure 3.11 (II) the bootstrap support for Microsporidia being the deepest divergence amongst the eukaryotes was generally about 1.2 times as large as that shown in figure 3.11 with the programming error. With the LogDet implemented in PAUP* support for "Microsporidia deepest" started at 54% with no sites treated as invariant, and rose steadily and smoothly to over 90%, before falling back slightly to 85% with all sites removed (so in the range of 20 to 30% of sites treated as invariant support was 80 to 92%). Conversely the support for *Giardia* started at 46% and dropped to 7% with 30% of all sites treated as invariant, then increased slightly to 11% with all unvaried sites treated as invariant. There was still no support for Microsporidia and *Giardia* together (we observed one tree in 1,000 with this feature). Support for hypothesis (III) showed no notable changes. There was a practically no change in the overwhelming support for the plant / fungi grouping in hypothesis (IV). The general trend of support for the fifth set of

hypotheses was similar. In the case of (VI) there was an slight drop in support for *Thermus* being deepest of the eubacteria, with nearly all constant sites removed. This drop became apparent at p_{inv} of about 0.25, but more notable at 30% of sites treated as invariant (when support had dropped to 95%), decreasing to 85% support with all unvaried sites removed. Support for the alternative hypotheses of "cyanobacteria deepest" only increased to 2% (as large sampling error seemed to be the probable cause for the drop in support for "*Thermus* deepest").

In conclusion, the invariant sites-LogDet transform has picked up at least one case (II)(the position of Microsporidia) in which failure to account for both invariant sites and base composition effects has hidden an important feature of evolution (namely the earliest eukaryotic divergence). What is also pleasing is that in direct comparison to the invariant sites Jukes-Cantor transforms, the invariant sites LogDet transforms showed little evidence of sampling variance being a noticeable concern except possibly with all unvaried sites removed. Indeed in these comparisons the LogDet transform was apparently more stable than the maximum likelihood 5 parameter distance in PHYLIP (support with this p_{inv} transformation is shown as the hollow symbols in figure 3.11). With all unvaried sites removed, the LogDet transform was generally showing clear signs of lower sampling error than the 5 parameter distance, which was something of a surprise. However it is apparent that sampling error of very large distance estimates, and secondarily some of the deepest nodes in the tree, is rapidly increasing when 30% or more of all constant sites are treated as invariant, and this needs to be a concern with all transformed distance methods. It is partly indicating just how diverged some of these sequences really are when unequal rates across sites are considered.

3.7.4 The overall invariant sites-LogDet "tree of life"

In figure 3.12 we show the weighted tree as estimated by the neighbor joining algorithm applied to the 16S-like rRNA sequences, after the invariant sites-LogDet transform.. The tree did not change between $p_{inv} = 0$ and $p_{inv} = 32\%$, while the branch lengths shown are for 20% of constant sites removed. The edge lengths remain quite similar to those inferred by other transformations. The edge lengths show the eukaryotes to be more divergent in sequence than the prokaryotes, while thermophiles appear to be especially slowly evolving. These distinct differences in edge length are unlikely to be a compositional bias due to the properties of the LogDet transform. Overall the branch lengths suggest that for much of the tree, total amounts of change are not in the region that suggests randomisation. While the edge lengths increase approximately 15% when 30% of sites are removed, the edge lengths shown in figure 3.12 seem to generally overestimate distances measured in the unweighted number of substitutions per site by about 15 to 30% (due to the weighting of base changes by the inverse frequency of bases at the time the change occurs and due to the invariant sites transform probably inflating many distances over that which might be inferred by a continuous distribution of rates). It would be surprising if many proteins can achieve such low rates when all sites with deletions are removed, the proportion of sites implied to be invariant are removed, and a LogDet transform is then applied. (Note: the position of the root shown in figure 2.12 is arbitrary, although some authors e.g. Sogin 1991, Jeffares *et al.* 1995, favour this location. If, alternatively, the root is on the edge

leading to eubacteria e.g. see Iwabe *et al.* 1989, then there would be clear evidence for an increase in eukaryotic evolutionary rates).

A common argument for keeping all sites in an analysis, is that that removing the more rapidly evolving sites (especially those in insertion deletion regions) will obscure the resolution of the finer features of the tree. Here however we seem to have kept very good resolution, down to known features as fine as the grouping of purple bacteria, halobacteria, the plants, and even resolving the five animals apparently correctly. One of the most labile sequences of all appears to be that of nematode which, as we see later, shows many homoplasies despite being a relatively shallow branching lineage amongst sequences with very equal base composition. These misleading changes offer a partial explanation of why it may be so difficult to resolve the early evolution of the metazoans. The relatively long edges leading to many of the early and middle eukaryotes suggest that this part of the tree may be prone to convergent and parallel changes. In contrast the eubacteria show much less divergence (only about 0.2 substitutions per site for many pairs), but may have diversified into the major groups in a short period of time, and it is possibly for this reason their resolution is proving difficult (e.g. see van De Peer *et al.* 1994).

One point of this tree that deserves mention is the ratio of the height of the branching point of the crown group (here plants, animal and fungi) to the deepest eukaryotic divergences. As Hasegawa *et al.* (1992) point out the date of this divergence is often inferred to be approximately 1.2 billion years ago. Extrapolating back on earlier trees inferred from 16S-like rRNA, the time of divergence of the earliest eukaryote (in most other trees taken to be *Giardia*) results in a date as early as 12 billion years, 3 times the age of the earth! (Hasegawa *et al.* 1992). In this tree, the same type of calibration (taking the average divergence of the crown group members, except the fast evolving nematode, and extrapolating back along the backbone of the tree) suggests the earliest eukaryotic divergence is 2 to 3 times as old as the crown group. This gives a much more credible age of 2.4 to 3.6 billion years ago. Given the base compositions of present day taxa (including dozens of early eukaryotes not shown in this tree), it seems possible that the "backbone" eukaryotic lineage had a near equifrequency base composition for the whole of this time. This, in turn, would make such estimates more valid, as would further confirmation of the quasi clock-like evolutionary rate of many eukaryotic lineages. It would be desirable to have better statistical tests of both these features in future.

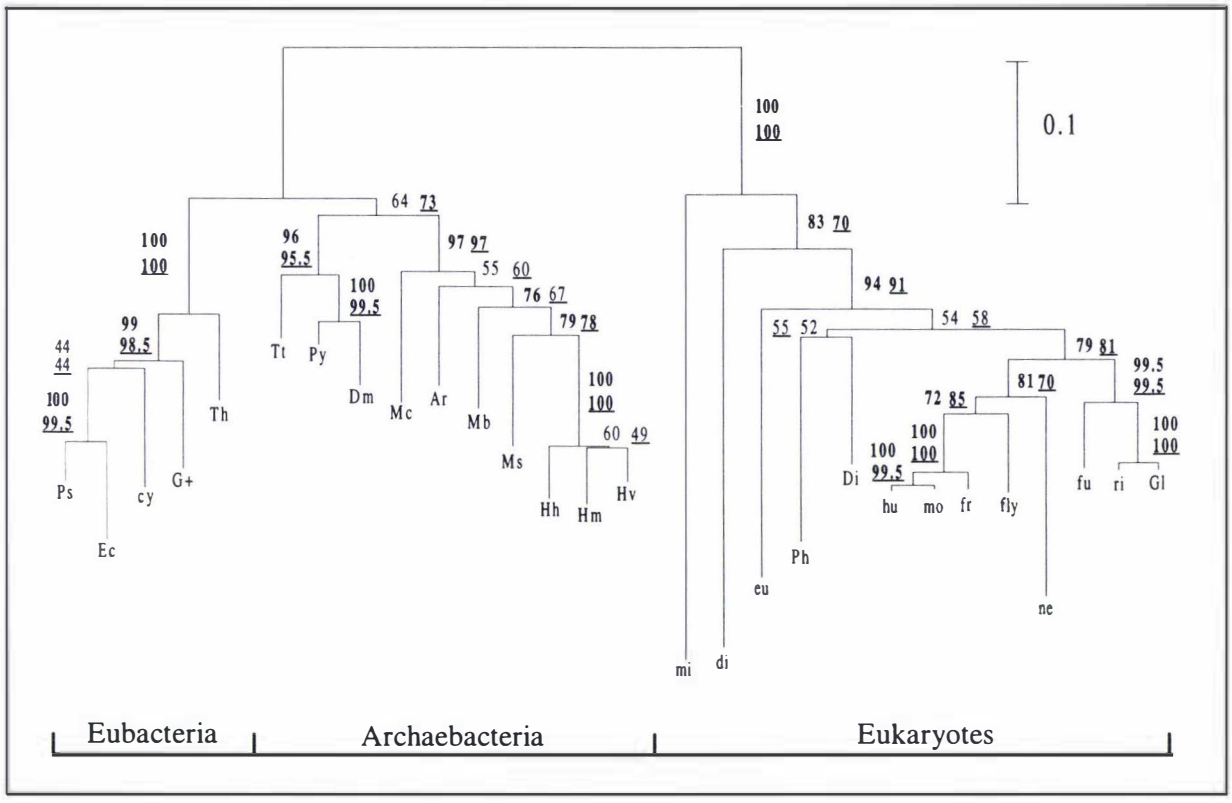


FIGURE 2.12 The unrooted weighted tree recovered by applying the neighbor joining algorithm to the small subunit rRNA alignment of Gouy and Li (1989a) after transformation by the LogDet method with 20% of unvaried sites removed (with sites removed in proportion to the frequency of nucleotides at constant sites). Numbers indicate the frequency with which an internal edge was recovered after applying the same method to 1000 bootstrap samples of the original sequences (values of 70% and above are in bold). (the underlined numbers are the frequencies obtained with the erroneous LogDet calculation in "Trees"). The edge lengths correspond to the LogDet distance (a weighted form of the number of substitutions per site, eq. 3.3.1-6). The sequences are from the following species (from left to right): The eubacteria, **Ps** = *Pseudomonas aeruginosa* (α subdivision of purple bacteria), **Ec** = *Escherichia coli* (α purple bacterium), **cy** = *Anacystis nidulans* (cyanobacterium), **G+** = *Bacillus subtilis* (low GC gram-positive bacterium), **Th** = *Thermus thermophillus* (a thermophilic bacterium, probably close to the green non-sulfur anaerobic photosynthetic bacteria). The archaebacteria, **Tt** = *Thermoproteus tenax*, **Py** = *Pyrodictium occultum*, **Dm** = *Desulfurococcus mobilis* (all three sulfur metabolising archaebacterial thermophiles, sometimes called eocytes), **Mc** = *Methanococcus vannielii*, **Mb** = *Methanobacterium formicicum*, **Ar** = *Archaeoglobus fulgidus*, **Ms** = *Methanospirillum hungatei*, **Hh** = *Halobacterium halobium*, **Hm** = *Halococcus morrhuae*, **Hv** = *Halobacterium volcanii* (all "methanogenic" archaebacteria, with the last three being extremely salt tolerant, while *Archeoglobus* is an thermophilic methanogen). The eukaryotes, **mi** = *Vairimorpha necatrix* (a microsporidian), **di** = *Giardia lamblia* (a diplomonad, a flagellated amitochondrial eukaryote), **eu** = *Crithidia fasciculata* (of the euglenoid group), **Ph** = *Physarum polycephalum*, **Di** = *Dictyostelium discoideum*, **hu** = *Homo sapiens*, **mo** = *Mus muscus*, **fr** = *Xenopus laevis* (frog), **fly** = *Drosophila melanogaster*, **ne** = *Caenorhabditis elegans* (nematode), **fu** = *Saccharomyces cerevisiae* (yeast), **ri** = *Oryza sativa* (rice) and **Gl** = *Glycine max* (soybean) (both angiosperms).

When taking out 28% or more of the constant sites we begin to encounter the occasional negative determinant (range 1 per 100 bootstrap samples, each of 378 unique distance values with 28% of constant sites removed, increasing to about 1 per 10 bootstrap samples, with all

constant sites removed). The program "Trees" (Penny *et al.* 1993) set all such values to 9.990, which is rather a large number (other bootstrapped LogDet distances for this data rarely exceeded 1.5, even with all constant sites removed). To consider how much effect these large 9.99 values were having upon the neighbor joining algorithm, and the corresponding bootstrap partition frequencies, they were reset to 0.8, 1.1, 2 and 3, prior to tree selection. Even for the bootstrapping of samples with all constant sites removed, these modifications lead to only a small increase (less than 1.5 %) in the bootstrap support for the favoured hypotheses in figure 3.11. Consequently we do not see the value assigned to a negative determinant as critical, the main problem must therefore reside in the increasing variance and bias in all the other distance estimates.

Sometimes an especially rapidly evolving species can produce extra noise in a phylogenetic analysis. Using PAUP 3.0, the nematode worm had the highest homoplasy index of any species (greatest number of convergences and parallelisms implied by parsimony reconstruction). These distance analyses were rerun with the sequence for nematode (*Caenorhabditis elegans*) removed. The results remained nearly identical to those presented in figures 3.11 and 3.12, except the stability of the grouping of the remaining animals increasing dramatically when all constant sites were included. More stability amongst the middle diverging eukaryotes was also observed, with the order (((crown group, *Dictyostelium*), *Physarum*), *Crithidia*) being clearly favoured over any others. This apparently increased resolution decayed away with the removal of constant sites, so that with all constant sites removed the analysis overall was very similar to those with the nematode included.

One approach to boosting the robustness of the invariant sites-LogDet transform to unequal rates across sites, is to remove sites from the data that are implicated to be the most rapidly evolving. To do this we imported the data into MacClade and then removed the approximately 5% of sites showing the greatest number of changes on the LogDet tree. The sum of sites showing 8 or more changes on this tree was 45 (equals 5.6% of the original 800 sites) and all these were removed. That these were more rapidly changing sites is supported by many distinct clusterings of sites with 8 or more changes along the sequence, which is consistent with local regions of higher rates substitution.

The same analyses performed earlier were repeated on the data after the above editing and after removing 20% of constant sites (from the original 800). The results are shown in figure 3.11 as the black symbols detached from the continuous lines (LogDet transformation only shown). Notice that the bootstrap support remains similar for many hypotheses; interestingly those for the fungi-plant grouping do not change perceptibly, suggesting that this grouping is not an aberrant effect of the more rapidly changing sites. In the case of the root of the eukaryotic tree, this editing increases support for the Microsporidia being deepest branching. Editing has also increased the support for *Physarum* being a deeper branching lineage than *Dictyostelium*. Another notable change is that *Archeoglobus* has more support for being deeper amongst the methane producing archaebacterial group, suggesting that the sites supporting its inclusion near the halobacteria are faster evolving than the average for this data. This does not necessarily

suggest that support from the more rapidly evolving sites is misleading (in this instance we suspect it may be useful to resolve some of the shallower branching), rather we seek just to diagnose the effect of these sites (which overall does not seem substantial). Very similar effects were observed with the Jukes-Cantor correction (results not shown).

As mentioned earlier we have considered how a variety of other commonly used distance corrections performed in these analyses (the Kimura 2P distance, with or without a gamma distribution of rates across sites setting shape parameters of 10 down to 0.6, with 1 being favoured by the analysis in table 3.4), and the 5 parameter distance derived from the substitution matrix used in DNAML (Phylip 3.5, Felsenstein 1993). All gave very similar results to the Jukes-Cantor distance, for example all strongly favouring *Giardia* as the deepest branching amongst the eukaryotes. The 5 parameter distance appeared to have the largest variance of the transformations used from PHYLIP (evidence for this included a more prominent mixture of unlikely phylogenetic groupings amongst the partitions not included in the bootstrap consensus tree). The bootstrap values when 20% of unvaried sites are removed then the 5 parameter distance estimated, are shown in figure 3.11 as the hollow black symbols. With PAUP* the Tamura-Nei distance appeared to have the worst sampling errors, which were distinctly worse than even the general time reversible distance (this comparison was made by noting how often these distance estimates had to be adjusted due to a sample inferring an infinite distance).

We have assisted Dr David Swofford to program the general time reversible distance with a distribution of rates across sites into the program PAUP*. The results of using this distance transformation (in a test versions of PAUP*) on this data when assuming a Γ distribution (and shape parameter $k = 1$) are very similar to the time reversible 5 parameter distance in Phylip with about 25% of invariant sites removed. Both methods tended to show lower bootstrap support then the simpler distances (e.g. the Jukes- Cantor and Kimura 2ST), especially when making modifications to allow for a realistic distribution of rates across sites (as estimated in section 3.6 for this data). The generalised time reversible distance with unequal rates across sites also gave high (approximately 80%) bootstrap support for *Giardia* deepest amongst the eukaryotes. Accordingly we suspect it is no more robust to nonstationary base composition than the simpler distances which also assume stationary base composition.

Lastly, to check that the neighbor joining method was not giving bootstrap support which was overly distorted by its very local tree search strategy, we employed extensive searches using the related minimum evolution criterion. With the original data this criterion gave the same optimal trees as neighbor joining, in spite of the quite thorough searching allowed by PAUP*. A bootstrap run of 1000 was performed using the minimum evolution, starting with the same pseudorandom number used to give the neighbor joining bootstrap supports in figure 2.12 (the number being 5). Selecting a TBR search for each replicate (e.g. see Swofford and Olsen 1990), between 10,000 and 40,000 minimum evolution trees were evaluated starting from the neighbor joining tree of each replicate. It was obvious watching the output, that 3 or 4 times during each replicate a better minimum evolution tree was found. However this difference in preference did not show up strongly in the overall bootstrap support. The resulting bootstrap consensus tree,

with bootstrap support given in percent, was (((((Ps, Ec):100, cy):39, G+):98, Th):100, (((Py, DM):100, Tt):**100**, ((((((Hm, Hv):60, Hh):100, Ms):**89**, Mb):**81**, Mc):**45**, Ar): **99**,):**76**, (((((((((hu, mo): 99.5, fr):100, fly):**97**, ((ri, Gl):100, fu): 99):65, ne):**47**, (di, Ph):47):**41**, eu):96, di):83, mi):100). We have highlighted the more significant differences from the neighbor joining analysis. In general there has been an increase in support for most of the already well supported edges in the tree. A minor rearrangement occurred in the archaebacteria tree (from both the optimal minimum evolution tree, and the consensus tree from bootstrapping neighbor joining). In either case bootstrap support was low at approximately 50%. The last notable feature was that the minimum evolution bootstrap replicates (but not the optimal minimum evolution tree) tended to place the nematode just outside the remaining crown group taxa, with bootstrap support of only 47%. Overall it would seem that in this instance the differences are minor and neighbor joining does not appear to be strongly biased by its tree search procedure, if anything it generally tends to underestimate well supported edges, compared to what a fairly extensive minimum evolution search gives. These findings are consistent with those in Charleston (1994), and Charleston *et al.* (1994) who showed very small differences in the errors made by neighbor joining on different tree topologies (distinct unweighted, unlabeled trees).

3.7.5 The relative performance of ML and parsimony methods on this data

The 5 parameter ML method in Phylip: Compositional bias was expected to be a prominent factor affecting these methods, while unequal rates across sites would certainly alter bootstrap support for specific hypotheses. The 5 parameter ML method of DNAML (Felsenstein 1993), produced trees similar to those produce by the Jukes- Cantor / neighbor joining combination. This method clearly favoured *Giardia* being deeper than the microsporidian sequence. However, careful analysis of 11 representative taxa using DNAML, showed this method to apparently be even more sensitive to anomalies in the data than any of the distance methods (the taxa were human, *Dictyostelium, Physarum, Crithidia, Giardia, Vairimorpha, Halobacterium, Desulfurococcus, Methano sp., Escherichia,* and *Thermus*). This reduced subset had the Trypanosome (*Crithidia* sp.) moving from its apparently correct position in the 28 taxa consensus tree (fig. 3.12), to a position next to human (this involved moving past two other branches, *Dictyostelium* and *Physarum*). The Kishino-Hasegawa test (as implemented in DNAML) verified that the log likelihood between this optimal ML tree and the LogDet neighbor joining tree was significant. An examination of figure 3.8 shows no evidence for base composition causing this problem, nor was it as evident with any of the distance methods used on the same data set (including the 5 parameter distance correction in DNADIST, Felsenstein 1993), although *Crithidia* sometimes moved closer to the "crown group" than *Physarum*, but not closer than *Dictyostelium*.

The same study was repeated after removing various proportions of the constant (taking account of their distinct base composition) followed by ML tree estimations were. By doing this we were effectively performing maximum likelihood tree estimation while taking account of the distinct base composition of the slowest evolving sites. Either doing this, or else modeling some sites as invariant but otherwise similar in base composition to the variable sites (and option in

DNAML 3.5, Felsenstein 1993), made little difference to tree selection in this instance. The log likelihood difference between the favoured ML tree and the LogDet tree remained nearly constant with 0 to 32% of sites treated as invariant. More recent runs with likelihood in PAUP* (Swofford 1995) tend to confirm that current homogeneous 4-state ML methods will not put Microsporidia as the sister taxa to the other eukaryotes (with these sequences), and are also sensitive to the misplacement of nematode. Results such as these caution us that while the ML tree selection criterion appears to have desirable robustness to violation of rates across sites (see chapter 5 of this thesis), and also base composition in simple models (Fukami-Kobayashi and Tateno, 1991), we should not assign it magical properties especially when the model is violated in multiple ways. Given results such as this we see no reason to believe that it will always do better than distance methods, which apparently have some interesting robustness properties of their own. (This claim is assuming we can be confident of the other biological evidence which shows *Dictyostelium* as the most closely related of the middle eukaryotes to the crown group).

Unweighted parsimony on the observed sequences: A tree was selected from the observed untransformed data with parsimony, and then subjected to a bootstrap analysis, the results of which are shown in figure 3.13. As with the 5 parameter ML method, this sequence pattern based method also moves *Crithidia* next to the crown group, and *Giardia* replaces *Vairimorpha* as the deepest branching eukaryote (the optimal 11 taxa tree selected by ML is a subtree of this tree). Other changes are *Archaeoglobus* moving to the deepest position amongst the methanobacteria, while nematode and *Drosophila* have come together. Bootstrap values have changed considerably in some parts of the tree, for example, there is apparently strong support for cyanobacteria branching more shallowly than the gram positives, while values supporting the structure in the middle and lower parts of the eukaryotic tree have dropped dramatically. We interpret this last effect as due to a fair amount of homoplasy amongst these eukaryotes.

Curiously, the Fitch-Margoliash method shows similar differences with respect to the LogDet tree. An examination shows that some of this methods differences from the LogDet tree branching order require only a 15% or so alteration in bootstrap support to become favoured, while others are more substantial requiring a > 25% change in support. To reiterate, neighbor joining using the observed distances has the group (plants, fungi and all animals except nematode), then a split with nematode, then a split with *Dictyostelium*, then a split with *Physarum*, then a split with *Crithidia*, then a split with *Giardia*, and so on. It certainly appears that the attraction (which must be due to convergences and parallelisms) between nematode and much earlier eukaryotes can cause problems in this part of the tree for most methods. Again this result is unexpected and presently undiagnosed, especially given the fairly even base composition in these taxa as seen in figure 3.8.

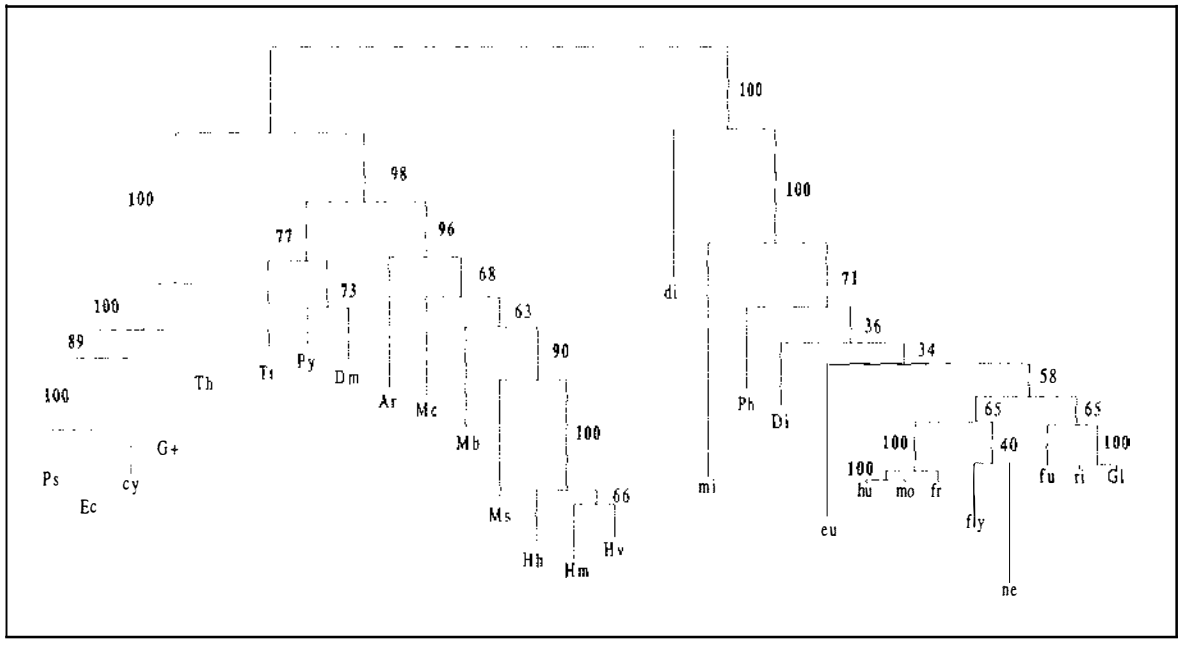


FIGURE 3.13. The unweighted maximum parsimony tree for the observed nucleotide patterns, along with bootstrap support for internal edges. Edge lengths are approximately to scale, and the taxa are as listed in the caption of fig. 3.12.

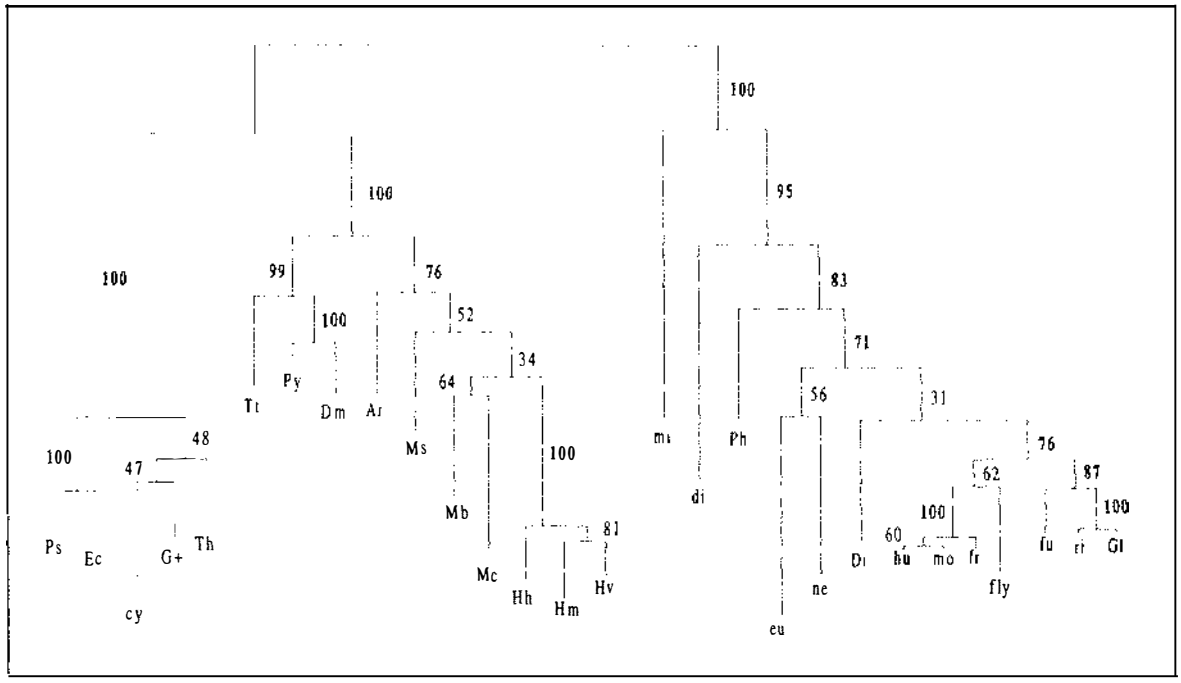


FIGURE 3.14. The transversion parsimony tree for the observed data recoded as purines and pyrimidines, along with bootstrap support for internal edges. Edge lengths are approximately to scale, and the taxa are as listed in the caption of figure 3.12.

3.7.6 Does an analysis of just transversional changes help?

When amounts of change are large, using only transversional changes can help by negating what are usually expected to be more common and misleading transitional changes. One way of achieving this is to map the states A and G to state R (purines) and map the states C and T to Y (pyrimidines), so any change R $\leftrightarrow$ Y must be a transversional change. An added rationale for this "mapping down to two states" is that transitional changes (A $\leftrightarrow$ G and C $\leftrightarrow$ T) are not only

proceeding at a higher rate then transversional changes, but are likely to be the prime proximal cause for changing "GC" content. As a result purine-pyrimidine content will remain stable longer e.g. sequences on the main diagonal of figure 3.6 will remain stable at 50% R / Y indefinitely. Consequently various groups (e.g. Weisburg *et al.* 1989) have advocated using transversional changes to analyse data where unequal base compositions are suspected of causing problems with tree inference. Tree inference based on just the R / Y recoded data for all 28 taxa was explored with a variety of tree estimation methods.

Firstly we applied unweighted maximum parsimony (MP) to the observed R / Y sequence patterns. The transversion MP tree (e.g. see Swofford and Olsen 1990), along with bootstrap proportions, is shown in figure 3.14. A most notable change is that while the *Crithidia* sequence has moved one step towards the crown group, the nematode sequence has jumped three internal edges to group with *Crithidia*. Further, if we examine the 1000 bootstrapped trees, not one of them has the animals grouped together! Another notable rearrangement is that there is no longer any clear support for *Thermus* being the deepest branching of the bacteria, while 23% of trees place the cyanobacteria deepest (vs 28% placing *Thermus* clearly deepest). Another notable change is that the methanomicrobials and the methanobacterials group together with moderate support (something not apparent in the other analyses, but suggested by other gene trees). *Archaeoglobus* is again the deepest of the methanogenic bacteria, while the methanomicrobial, halobacterial grouping now has only 20% support, having been quite well supported in all the earlier analyses using four states. Lastly, there is strong support (95%) for the microsporidian sequence diverging earlier than *Giardia*. Of these rearrangements one definitely contradicts biological knowledge (nematode position), another is quite unlike most published rRNA eubacterial phylogenies (e.g. that of Olsen and Woese 1993), another strongly supports a novel feature of the invariant sites LogDet sequence analysis ("Microsporidia deepest"), while another disagrees with our previous analyses of these sequences, but agrees with trees from other genes (the methanococcus / methanobacteria grouping).

The data were also converted to distances, followed by Poisson model corrections then tree estimation with the neighbor joining algorithm and minimum evolution. Corrections included allowing sites to follow a gamma distribution with shape parameter $k = 1$. These corrections were made with the program PAUP* by running Kimura 2ST distances, but having just the transversional changes counted. All the trees obtained by this general procedure were practically identical, being similar to both the LogDet tree, and the tree of figure 3.14. They generally gave similar amounts of bootstrap support to the LogDet tree (e.g. "Microsporidia deepest" has low 90's support, plants with fungi 88%). The support for *Thermus* at the root of the eubacteria was however low, at 66%. There were also some surprising rearrangements, the eocyte tree was favoured over the archaebacteria tree and the halobacteria tree by 50: 18: 10 (respectively). The middle eukaryotes showed yet another rearrangement with the order *Dictyostelium*, *Crithidia*, nematode, *Physarum*, and *Giardia* as progressively more deeply branching lineages. These arrangements from the uncorrected transversion parsimony tree of figure 3.14 are relatively minor (being in regions with low bootstrap support), except for splitting up the archaebacteria

which was a complete surprise. Just why the corrected transversion distances are doing this is unknown, although it could possibly have something to do with tied neighbor joining trees which occurred only when bootstrapping this data. Clearly, correcting for multiple hits in the transversional changes is making at least as much difference to the inferred trees as corrections with 4-state data. Thus it should not be acceptable to quote just the transversion parsimony tree estimated from the observed data, in this type of analysis. Overall then using just transversional changes has been insightful, but the number of rearrangements in the trees seen when they alone are used is quite surprising, and suggests the need for closer study of the methods used with this type of mapped down data.

3.7.7 The validity of grouping transversions

An interesting point regarding analysing just transversions relates to the grouping of states in a Markov process. An assumption of current tree building methods is that the probability of a particular site substitution on one edge *m* in the tree, does not depend upon evolution occurring in another edge in the tree, unless that edge is between *m* and *r*, the tree's root (e.g. Steel *et al.* 1993a). While a 4-state i.i.d. Markov process on a tree will meet this condition, grouping states can violate the assumption of independence, in just the same way that grouping states can invalidate the probabilities obtained from a Markov chain (e.g. proposition 5.9 of Iosifescu 1980). Characterise the transition matrix on each edge of the tree as, $\mathbf{P} = \begin{bmatrix} * & a & b & c \\ d & * & e & f \\ g & h & * & i \\ j & k & l & * \end{bmatrix}$.

Then states 1 and 2 vs 3 and 4 are groupable without distortion of the model if, b + c = e + f, and g + h = j + k (giving **P** at most 10 free parameters, since * is one minus the other entries in that row). Interestingly evolution by the generalised Kimura 3ST model will always meet this condition, but a stationary time reversible model need not. The condition for evolutionary parsimony to be consistent is similar, except we replace the constraint on sums of entries with b = c, e = f, g = h, and j = k (known as the balanced transversion model, e.g. Navidi and Beckett-Lemus 1992). Consequently, while both procedures place emphasis upon transversions, they are guaranteed to be consistent under slightly different models. There is evidence that evolutionary parsimony is quite robust to violations of its constraints on **P** (Navidi and Beckett-Lemus 1992), and we suspect robustness when analysing just transversional changes, but this remains an important area for study.

Table 3.7 Unambiguous substitutions on the invariant sites-LogDet N.J. tree

	To *Giardia*					To *Vairimorpha*			
	A	G	C	T		A	G	C	T
A	-	13	17	1	A	-	0	1	10
G	1	-	10	1	G	35	-	3	28
C	1	8	-	0	C	18	5	-	15
T	3	5	27	-	T	5	1	1	-

Table 3.7 shows the numbers of unambiguous substitutions on the edges leading to *Giardia* and *Vairimorpha*, based on the LogDet-neighbor joining tree. Assuming the equi-frequency base composition of the sequences prior to evolution down these edges, we can make a Pearson X^2 test of the assumptions for grouping states into transitions and transversions. We find no significant difference in the observed frequency of A → C + A → T vs G → C + G → T substitutions, or C → A + C → G vs T → A + T → G changes on the edge leading to *Giardia*, but do reject these equalities in the case of *Vairimorpha* ($P < 0.001$). The ideal test that the transversional changes were groupable would be to compare the likelihoods of the data with and without the transition matrix constraints in place. Then of course the likelihood model must entail allowing for nonstationary base compositions, plus distinct base compositions in different rate classes, to make the test valid. Again it is noticeable just how oppositely skewed the substitution process in these two lineages is, with the expectation from the very rare singleton site changes that it has been this way for a very long time.

3.8 CHECKING THE INVARIANT SITES-LOGDET TREE RESULTS

In this section we aim to put the results of the previous sections into some sort of perspective by critically looking at the support for our hypothesis. It is very important to be eclectic (to gather together the best from diverse sources) when evaluating evidence for evolutionary relationships using anciently diverged molecules. This is because it is certain that the real process of evolution is not i.i.d., nor are the fixed informative changes neutral, in the sense that their persistence must be due to their being selectively favoured over alternatives (although the original substitution may have occurred at a time when it was effectively neutral).

3.8.1 Using just the most conserved informative sites to avoid model uncertainties

In this section we seek to test the validity of the inferences made earlier using i.i.d. model dependent methods, when we know that the model is almost certainly a covarion type model of evolution (see the discussion section of chapter 2). The method used here harks back to those of traditional systematists (e.g. Hennig 1966) who strove to find a set of unambiguous characters that clearly demarcated phylogenetic events. Hennig (1966) called these characters synapomorphies, and desired to polarize all of them, but there is nothing to prevent their generalisation to unrooted trees. Here we define a set of unrooted characters which would unambiguously support our six prespecified hypotheses I-VI, and call them "perfect" characters.

While our data set contains sites with widely varying rates, one would hope that at least some of these hundreds of sites have evolved at a sufficiently slow rate that the partitions they define are in perfect agreement with partitions in our inferred phylogeny. Thus we want to identify very slowly evolving sites, and count how many of them support each alternative to our prespecified hypotheses. Three rather strict criteria can be used to identify very slowly evolving but potentially informative sites. The first criterion, defines type (1) sites, which have just two states in the whole data set, and the partition defined by this matches one of the hypothesised

partitions in I-VI exactly. The second criterion for type (2) sites, is a slight relaxation of the first; a site may have more than two states, but must fit one of the six sets of prespecified tree partitions perfectly, with no possibility of contradicting any other partition in the tree. This eliminates sites contributing to the resolution of more than one of the possible partitions. Our definition of a site to meet this criteria is a site like that which meets criterion (1) but is allowed up to two singleton states (had by only one taxon) which occur in taxa well away from the "informative" partition that site supports (to avoid ambiguity). Finally, type (3) sites are defined as for type (1) sites, except characters are recoded as either purines (R) or pyrimidines (Y), so these are sites with apparently unique transversional changes supporting just one resolution of the hypotheses.

A similar type of analysis was used to evaluate the support for the archaebacteria as monophyletic in Olsen (1987). His criterion for accepting sites was more relaxed than those used here (he allowed one taxon per group to have a different state to the others). This criteria could lead to some characters showing quite ambiguous and perhaps contradictory support for the group they are claimed to support. Our criteria do not have this problem. They also make it more straightforward to evaluate multiple hypotheses on the same data set, since a site can only support one hypothesis. The results of this analysis are shown in table 3.8.

The "perfect sites" counting is focused upon just those hypotheses where one alternative received at least 90% bootstrap support with some number of invariant sites removed. The archaebacterial hypothesis is also included, since it had received very high support with no constant sites removed (figure 3.11, Gouy and Li 1989a). It is possible that the loss of support for the monophyly of archaebacteria is an artifact due to an "anti-Felsenstein zone" effect (mentioned previously in section 3.4.3 and covered more completely in chapter 5), which could be triggered by the bias in distance transformation (due to either stochastic error, or a systematic error of the type discussed in section 3.4). Thus, in a sense, the "perfect sites" counts are a test of not only the support for monophyly, but of the validity of the tree building method that was applied to the bootstrapped data.

The results of our "perfect sites" counting are shown in table 3.8. In the case of the first set of hypotheses, there is plenty of support for the monophyly of the archaebacteria. In the case of character types (1) and (2), there is an 8: 1: 2 advantage for the archaebacterial tree over the eocyte and halobacterial tree hypotheses respectively. There is little evidence of excess parallel changes in this vastly reduced dataset, and there is no evidence for the eocyte tree being better supported than the halobacterial tree (something also evident in the bootstrap analysis). It is also interesting to note that many of the "perfect sites" involve transitional changes (specifically 6 of the 8 sites of type 1 and 2). If we add up all three types of "perfect sites", then the archaebacterial tree is favoured by 12: 1: 2. We do not anticipate that there is a serious alignment bias in this data, as Gouy and Li (1989a), used secondary structure to check the alignment and all regions of ambiguity of alignment were excluded. It is desirable in future to check if any of these "perfect sites" are complementary base pairs changing together in which case we may wish to exclude one of the pair. We suspect that perhaps two of them are, as Olsen (1987) noted that

of the 11 conserved sites (by his definition) supporting the archaebacteria as monophyletic, there were two pairs of sites (hence 9 that are apparently independent). This part of the analysis agrees very well with the studies of Olsen (1987), and Olsen and Woese (1989).

"Perfect sites" of type (1) and (2) give a tie in terms of the deepest eukaryotic branching, with 2 sites supporting each alternative resolution. However, amongst transversional "perfect sites" type (3), there is clear support for the "Microsporidia deepest" hypothesis (with 8: 2: 1 for the alternatives defined in table 3.8). One explanation for this difference is that transitional changes, especially on the long external edges leading to *Giardia* and the microsporidian, have erased much of the original signal. This explanation however deserves further scrutiny. If later transitional changes were erasing this signal, then why did this same factor not erase many of the type 1 and 2 changes supporting the archaebacterial tree? The best explanation seems to be that a covarion model was operating and that many of the archaebacterial sites were no longer labile to substitution on the early eukaryote, or any other lineages. Another explanation is that the bias is simply a sampling effect (although this does not seem probabilistically likely given the quite different numbers of each type of site). If the covarion explanation is accepted, it suggests that these sites which changed early in eukaryote history were still labile to transitional but not transversional changes for some time later. Over all types of sites, the Microsporidia first hypothesis has 9: 4: 3 support which is still reasonable (in chapter 6 we look at tests for the significance of such a result).

Table 3.8 "Perfect characters" for hypotheses which received > 90% bootstrap support.

Hypothesis 1.	Archaebacterial tree	Eocyte tree	Halobacterial tree
character type 1	75, 98, 145, 329, 505	506	
character type 2	146, 161, 176		108, 128
character type 3	75, 98, 112, 125, 196, 355		108, 128
Hypothesis 2.	Microsporidia first	*Giardia* first	Micro + *Giardia*
character type 1	190, 251	331, 504	334
character type 2			64
character type 3	39, 43, 65, 190, 338, 393, 455, 533	197, 454	477
Hypothesis 4.	Plants + Fungi	Animals + Fungi	Plants + Animals
character type 1			
character type 2			
character type 3			
Hypothesis 6.	Thermus first	Cyanobacteria first	Thermus + Cyanobact.
character type 1	374, 422		
character type 2			
character type 3	228		

Note: Numbers refer to the site index for the reordered dataset of Gouy and Li (1989a) (the original data set minus all sites showing one or more missing characters, with all constant sites then removed).

There are no "perfect sites" supporting the resolution of the animals, plants and fungi. This could be due to a short period of time, and no major functional changes. It does cast some doubt on the reliability of the bootstrap support for this resolution. However there are sites which apparently change only once amongst the sampled eukaryotes that do support this hypothesis and importantly, no such sites support the alternative resolutions.

Support for *Thermus* being the earliest branching of the eubacteria we studied is spare (3 sites), but importantly there is no contradiction to this support. We take this as being in favour of interpreting the bootstrap support for *Thermus* first as likely to be well founded (for this data set).

It is interesting that there are many more "perfect sites" supporting archaebacterial monophyly than the other hypotheses. This may also suggest a considerable period separating these from all other taxa. Alternatively perhaps it represents a period of positve selection for functional substitutions which are of such importance that they are fixed in all the descendants studied (and that the rate of these changes was elevated relative to majority of the substitutions which the tree building algorithms use to estimate edge length).

Also surprising are just how many of the perfect sites are transitional changes and not the more conservative transversional changes. A count can be made by noting that any transversional changes defining "perfect sites" of type 1 or 2 will show up in "perfect sites" of type 3. For the favoured hypotheses (the first column in table 3.8) only 3 out of 12 substitutions of types 1 and 2 are transversions, with just 2 out of 7 for the alternative hypotheses (a test of these different proportions is clearly highly significant over the expected value if the "perfect site" substitutions were randomly chosen from all 12 possible changes, only 4 of which are transitions). This tends to suggest that the model is not i.i.d. (since the expectation should then be transversions predominant amongst the variable sites showing the fewest changes). An alternative interpretation is that rather common, near neutral, transitional changes are being subsequently frozen by what is probably strong stabilizing (or purifying) selection. This is a form of covarion model, and this perhaps should be considered one of its logical predictions if the form of model proposed by Fitch and Markowitz (1970) is operating (a prediction I have not seen noted elsewhere). Another alternative, which seems more ad hoc, is that these were always selectively advantageous changes, and they just happen to be mostly transitions.

Amongst other things this finding would tend to suggest that transversion parsimony may be excluding a lot of the most important information when studying anciently diverged molecules. This in turn may help to explain the somewhat erratic behaviour of all the transversion based methods studied in section 3.7.6. One way to test this hypothesis is to suggest that the transitions should define most deeper branches in the tree as well as the transversional substitutions. We have made this comparison first removing 20% of the unvaried sites by method (3) (in proportion to their frequency in the constant sites), then estimated a corrected distance matrix from the transversions, and separately the corrected transitional rates implied by Kimura's (1980) formulae (e.g. see appendix 2.6 for the same sort of methods used on all the pathsets). To

indicate the degree of difference, we bootstrapped the data 1000 times, and built neighbor joining trees for transversion, then transition distances (using the program PAUP*, Swofford 1995).

The transversion tree was very similar to that described already at the end of section 3.7.6 which was built with transversional changes corrected with a gamma distribution of rates across sites. The transition distance tree was also very similar to the LogDet tree. The notable differences between the transition and the transversion trees were: Stronger favouritism (80's b.s. support vs 50's with tv's) for Thermus deepest, then Bacillus second deepest in eubacteria. Much stronger support for the archaebacteria grouping (Ms, (halobacteria)) (89% vs 38% with tv), but lower support for the grouping (Dm, Py) (57% vs 100% with tv). The archaebacteria monophyletic had support of 72% with transitions vs 49% with tv. The most marked changes were amongst the eukaryotes with: 100% transition b.s. support for *Giardia* at the root, and very equivocal whether the microsporidian or *Crithidia* was next deepest (consensus tree had *Crithidia* deeper but only in 35% of replicates). The slime molds, *Dictyostelium* and *Physarum* were of uncertain affinity, with very low b.s. support (33%) placing them as sister taxa to the still strong grouping of plants and fungi. The animals remained monophyletic (68%) in strong contrast to the transversions, while nematode was favoured slightly as a sister taxa to the fly (58%). Overall two points are apparent: (1) generally the transitions appear as reliable as the transversions across this whole data set, (2) where the transitions were the most frequent perfect sites, the overall transitions distances tended to reflect this, and visa versa for transversions. Thus the alternating nature of transitions or transversions supporting different parts of the tree at different levels is upheld, as is the generally equally useful nature of the transitional and transversional changes.

Statistical testing of phylogenies based on "perfect sites" counts is possible assuming independence of sites and using multinomial probabilities. We describe such tests further in chapter 6, using these "perfect sites" counts as examples.

In conclusion to this section, it is important to consider what the rarest changes in the data are suggesting and to examine whether these changes are strongly supporting, mildly supporting, or even contradicting the results of other tree building exercises. In this instance there were no contradictions to the best supported partitions in the invariant sites-LogDet tree, but the conserved sites certainly suggested that support for the archaebacteria as monophyletic may an underestimate. On a data set which has no such "perfect sites" or sites nearly as conservative in their change, then we are putting an awful lot of trust into tree building methods which have not been evaluated under realistic evolutionary models (and to some degree all i.i.d. models must be suspect). The only conclusion can be that there will be some success, but inevitably some spectacular failures. The strong claim that diplomonads are the deepest branch in the eukaryotes which was started by just such an analysis in Sogin *et al.* (1989) and followed with great enthusiasm (examples are given in Siddall *et al.* 1992), is a probable example.

3.8.2 What level of bootstrap support is significant on our tree?

A critical question frustrating nearly every one doing phylogenetic analysis, is what level of bootstrap support should an edge receive in order to have X% confidence that partition in the tree is correct? (This assumes of course confidence that the assumptions of the tree estimation method are well met). When a hypothesis is specified *a priori*, the bootstrap support tends to be a conservative estimate of the true support for a partition (see Rodrigo 1993, Hillis and Bull 1993). If a partition is deep in the tree, some simulations indicate that a 70% bootstrap support can indicate up to 95% confidence. Here confidence is defined as,

1-{the probability of a type one error},

where a type one error equals the probability of this much support for an edge, when it is not in the true tree (e.g. see Felsenstein and Kishino 1993). Of course if we do not have much confidence in our methods of building trees being reliable, and expect it could be strongly biased away from getting the correct tree, then bootstrap support is not necessarily related to confidence of an edge being correct. Rather it simply gauges whether it is expected this edge will remain in the optimal tree as longer equivalent sequences become available (where equivalent means evolved by substantially the same process) (e.g. Felsenstein 1985, Penny and Hendy 1985).

A statement to the effect that under model simulations the bootstrap tends to be conservative, often acts as an open invitation for researchers to attach as much significance to their finding as possible. One such example occurs in Baldauf and Palmer (1993) who seemed happy to take 70% bootstrap support as 95% confidence that they had found a strong evidence for grouping animals and fungi together, never mind the fact that other parts of their tree looked decidedly wrong but also had similar bootstrap support! The issue of the calibration of the bootstrap is discussed more fully in chapter 6. For the moment ponder the consequences of the fact that on the tree in figure 3.12, some partitions with very good support will appear in almost all the bootstrapped trees. Four good examples are the partition of eukaryotes from all others, eubacteria from all others, the "eocyte group" from all others, and lastly the "methanogens" from all others. Trapped between these 4 solid partitions is the edge answering the question, are the archaebacteria a single group? Thus it is reasonable to expect that in most trees reconstructed from this data, there are only two likely alternatives to the archaebacteria forming a single group, and these we have named the eocyte tree edge and the halobacterial tree edge. In section 4.9, we see that tree selection in such a case reduces to selecting the largest value in a three dimensional multinomial space. Consequently there is limited opportunity for conservatism of a test statistic based on resampling, but also the real possibility of a Bonferroni type problem if we do not prespecify the hypothesis we are testing. This problem is examined in more detail in chapter 6, but for the moment note that confidently resolving the archaebacterial question, will probably require at least 80% bootstrap support to achieve 95% confidence. This is all the more obvious when it is remembered that there are really only three alternatives, and two of these can readily be claimed to be "a priori" hypotheses.

3.8.3 Other sequences supporting Microsporidia as earliest diverging eukaryotes

It is interesting to consider in more detail the probability that Microsporidia are indeed the most anciently branching eukaryotes. When the high bootstrap support for this hypothesis under the invariant sites-LogDet transform was determined two years ago, we were still suspicious and sought all possible ways to corroborate the finding. During a research fellowship at the Smithsonian Institution, I was fortunate to visit Woods Hole and discuss my findings with Dr G. Hinkle, a postdoctoral researcher in Dr Mitch Sogin's lab. One of the data sets made available to those attending the NSF Molecular Evolution course organised by Dr Sogin, was a combination of 16S-like, partial 23S-like and near complete elongation factor-G gene sequences of early eukaryotes, with outgroups amongst the archaebacteria and eubacteria. Greg Hinkle claimed that this combined data set still gave substantial support for *Giardia* being deepest (approximately 70% bootstrap support with the Jukes-Cantor correction, followed by tree building with the Fitch Margoliash criterion). The reliability of this result is in question given the base composition biases and prior analyses with invariant sites LogDet.

Accordingly the previous data set was divided, and the elongation factor sequences analysed alone (this data set was similar to that of Hashimoto *et al.* 1994, but included unpublished sequences from Dr Doolittle's Nova Scotia laboratory for a microsporidian and a parabasalid). With a standard method (Jukes-Cantor transform, then Fitch-Margoliash tree selection as implemented in "Fitch", Felsenstein 1993) these sequences gave 70% bootstrap support to the Microsporidia being deepest. Congruence is often a good indicator of reliable results, and it will be useful to analyse this data set again when it is made publicly available. However obtaining clear congruence between different genes with very deep divergences is not that frequent, and unpublished analyses with Peter Lockhart and Tony Larkum suggest that despite previous claims eubacterial elongation factor α and G genes and RNA polymerase genes, do not show clear congruence with 16S-like rRNA sequences when invariant sites are removed and base composition taken into account (and contrary to earlier claims, e.g.). Cautioning us in the case of the early eukaryotes, the relatively long and apparently conserved transfer RNA synthase genes which Brown and Doolittle (1995) analysed, give no clear resolution to a trichotomy of human, fungi and a microsporidian (bootstrap supports of very near 50%, with one standard procedure putting the microsporidian and fungi together!). It is important to strive to understand why "reliable" genes (such as α and β tubulins and elongation factors), and "sophisticated" tree estimation methods including likelihood, give apparently contradictory trees with other not quite so ancient eukaryotic divergences. (As an example see the trees reported in Baldauf and Palmer 1993, where the lack of congruence is striking).

Unfortunately there is a tendency to attempt to answer the most difficult questions in phylogenetics, with the most incomplete data sets. In the case of resolving early eukaryote sequences, there is a potentially very useful data set which is lame due to it lacking one sequence. 23S-like rRNA sequences are completely sequenced for four *Giardia* species, a parabasalid (trichomonad), *Physarum*, *Dictyostelium*, *Crithidia*, and a host of other "middle eukaryotes," but presently only 500 base pairs of the 5' region are sequenced from

microsporidians (and much of this is from a variable region not appropriate to very deep phylogenetic analyses, Vossbrinck *et al.* 1993, Baker *et al.* 1994). A complete 2kb microsporidian sequence added to this data set should approximately double the resolving power of the rRNA sequences. This would still leave open the question of which models are reliable for evaluating such early divergences, although a "perfect sites" analysis could be particularly informative.

3.8.4 Results of the application of split decomposition to this data

In this section split decomposition is used to further examine the relationships implied by our transformed distances and also properties of the distance transformations themselves. Split decomposition is a method for visualising non-tree biases in distance data (Bandelt and Dress 1992). For our purpose, all we need know is that it measures the extent to which the Buneman four point metric is violated. With 4 taxa if the Buneman four point metric is met then a resolved internal edge is a one dimensional link, with length proportional to some measure of separation of two groups. If the four point metric is not met then usually two sums of disjoint distances (hypothetically $d_{13} + d_{24}$, and $d_{14} + d_{23}$) will be greater than the minimal sum $d_{12} + d_{34}$, and in this case let $d_{14} + d_{23}$ be the maximal sum (in the extreme case two sums are equal and minimal, yet this can be treated in the same way). A weight of $1/2\ (d_{14} + d_{23} - d_{12} + d_{34})$ is assigned to the main edge which separates taxa 12 from 34, then a weight of $1/2\ (d_{14} + d_{23} - d_{13} + d_{24})$ is assigned to the edge which separates taxa 13 from 24 (this quantity is also called a "quartet" measure). This relationship with two orthogonal edges can be shown as a graph with rectangle, instead of a single edge, with the four taxa one at each corner. The more square the internal portion of such a graph becomes, the more the 4-point condition is apparently violated (and in the extreme case we have a square when two sums of disjoint distances are equal, and the third is dramatically larger). The method of Bandelt and Dress (1992) is a way of measuring this "boxiness" of all subsets of 4 sequences, followed by projecting as much of this information as possible onto a planar graph of all the species. For these calculations we used the "Splits Tree" program of R. Wetzel and D. Huson (available from D. Huson, Dept. of Mathematics, University of Bielefeld, email: huson@mathematik.unibielefeld.de).

Generally, split decomposition works best with a few taxa, since it is rare to see systematic biases in a large set of taxa all in the same "direction" so they can be shown on one planar graph. However application to our data set of 28 taxa revealed two consistent trends with all taxa included. Firstly, using Hamming (observed) distances gave the result shown in figure 3.15. Note the nearly square box with *Vairimorpha*, prokaryotes, *Giardia* and all other eukaryotes in this order at each corner (the fit statistic was 63.4%, while the split prime residue was 34.2%, see Bandelt and Dress 1992 for descriptions of these values). This pattern appeared to be indicate *Giardia* being drawn towards the prokaryotes due to their generally higher GC base composition, while *Vairimorpha* is partitioned with them by an orthogonal signal, due to what we expect is an historically correct signal. Elsewhere in the graph are smaller boxes indicating that the nematode sequence has attraction both to other animals and also quite strongly in the direction of the *Crithidia* sequence, which itself sits in one corner (along with the edge to

Physarum and the edge leading to *Giardia*, *Vairimorpha* and all the prokaryotes). All other resolved details were in agreement with edges of the LogDet tree which received substantial bootstrap support, except for the large polytomies in the archaebacteria and eubacterial lineages indicating possibly quite strong conflicting biases. Within the archaebacteria, this lack of resolution could be partly explained by *Archaeoglobus* being attracted to both eocytes and methanogens combined with not being specifically placed within the methanogens.

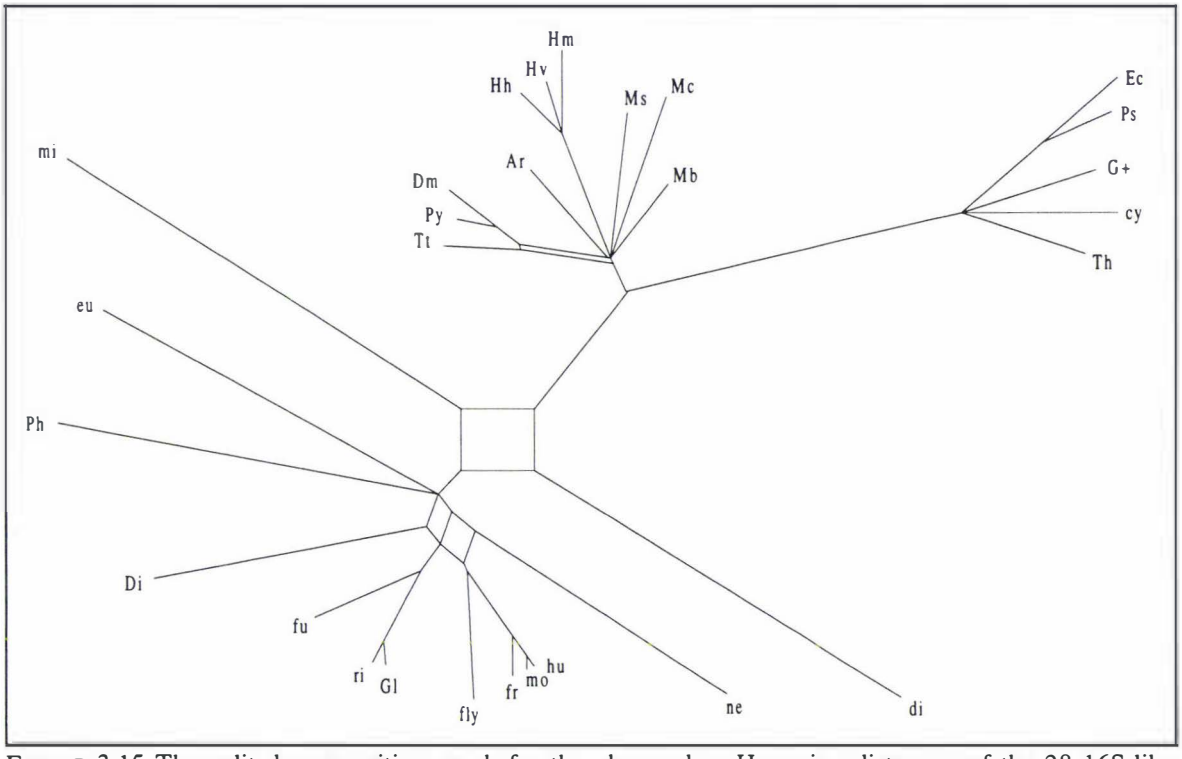


FIGURE 3.15 The split decomposition graph for the observed or Hamming distances of the 28 16S-like rRNA sequences (to scale except that the external edges to mi and di have been shrunk by 2/3rds). The two main areas of "boxiness" are around the resolution of the earliest split in the eukaryotes, and the divided affinities of the nematode (ne) sequence for other eukaryotes. There is a third fine "boxiness" indicating that *Thermoproteus* (Tt) is slightly less attracted to the methanogens than the other two eocytes. The unresolved portions indicate conflicting biases. The sequences, and sequence labels are as given in figure 3.12.

Running the same analysis but applying the Jukes-Cantor logarithmic transform (results not shown), the box with *Giardia* etc. around it becomes more rectangular (edges in a 2:1 ratio), with the *Giardia* now showing a closer relationship to prokaryotes, but the alternative of "Microsporidia deepest" is still visible (about 50% of the size of the alternative resolution). Removal of 25% of constant sites, sees the rectangle become much more extreme with the Microsporidia-prokaryote edge now only 1/4 the size of the alternative *Giardia*-prokaryote edge. These findings agree with our earlier claims that it is not just base composition similarity which causes attraction of edges in trees. Equally important is the fact that standard types of logarithmic distance correction dramatically overestimate the distance between sequences of unlike base composition (and so repel certain groups e.g. this Microsporidia sequence from prokaryotes). As would be predicted by this hypothesis, the application of a more complex

stationary distance correction formulae, here the K 3ST distance, made the situation slightly more extreme (i.e. more repulsion, especially with invariant sites removed).

Transforming the data with the LogDet saw the opposite trend. With all sites included, the edge separating *Giardia* from the other eukaryotes (Microsporidia excluded) decreased to only a 1/4 the size of the edge grouping Microsporidia with the eukaryotes. Removal of 25% of constant sites, saw support for "*Giardia* deepest" disappear from the graph of all 28 taxa. Application of the LogDet transform also removed the "boxiness" in the earlier graph indicating an attraction between nematode and the middle eukaryotes.

Recoding the data to purines and pyrimidines, we observed no boxes in the graph at all and a similar degree of resolution to Hamming distances for the 4-state data. *Vairimorpha* was partitioned with the prokaryotes by a larger edge than with the 4-state LogDet transform, and the nematode had left the animals completely and sat in an unresolved polytomy of the mid-branching eukaryotes, *Crithidia*, *Physarum* and *Dictyostelium*. These results confirm that transversional changes do tend to strongly support "*Vairimorpha* first", while transversional changes also strongly support nematode leaving the animals for the mid-branching eukaryotes. Future studies will include examining the graph when making 2-state distance corrections, including the LogDet, to this R / Y data.

It is important to evaluate the LogDet distance transformation on smaller sets of taxa, since split decomposition only shows boxiness if a bias in the Buneman four point metric is consistent across all taxa. With distant sequences, or just large numbers of taxa in a study, sampling error or uneven biases result in the split decomposition graph not revealing important features. If there is ambiguity in the ordering of the size of the three sums of distances between certain groups of taxa, then split decomposition will show a resolved edge if a certain grouping of all taxa indicates the same smallest sum of pairwise distances, while the other two distance sums fluctuate in rank size. Alternatively, if there is no consistently minimal sum of distances, then the relationship between those groups of taxa will be unresolved.

Of particular interest is whether the invariant sites-LogDet transformation has fully accounted for base composition biases, or perhaps even over-compensated. In fact, when we took high sets of just four or eight taxa, including *Giardia*, *Vairimorpha*, other eukaryotes and prokaryotes, then there was frequently still evidence of an attraction between *Giardia* and those prokaryotes with the highest GC content. Thus we suspect that in the most extreme parts of the tree, the LogDet is still undercompensating for base composition biases. This is most likely due to the confounding factor of rates across sites being only partly negated by the removal of constant sites. Accordingly, we expect similar biases to still exist with other methods which can accommodate base composition in other circumstances (e.g. the FD tests of Steel *et al.* 1993). The same deficiency would thus be expected in i.i.d. maximum likelihood tree selection methods which allow for unequal base composition (e.g. the 12 parameter per edge model of Barry and Hartigan 1987b) but do not separate sites into rate classes. The ML methods based on a Γ

distribution of rates across sites developed in chapter 5 do not separate sites by rate, so they too would be vulnerable.

On these smaller set of taxa, the nematode continued to show signs of attraction for particular middle eukaryotes including *Crithidia.* However here it also appears that the nematode is especially attracted towards *Vairimorpha*. This had not been so apparent in the tree analyses, but was hinted at by partitions including *Crithidia*, *Vairimorpha* and nematode appearing in a few of the bootstrap replications. While the particular nematode *C. elegans* itself is not an internal parasite, it is deeply nested within a group of parasitic organisms, and its ancestors were probably parasitic. This apparent attraction of just these three parasites which invade host tissues deserves further study beyond the scope of this thesis (*Giardia* and its relatives are, in contrast, confined to the gut, and so are not directly exposed to the immune system).

Overall, split decomposition has proven a valuable tool to help reinforce our diagnoses of how the LogDet transform and other distance corrections are interpreting this data. The fact that a graph can lose a lot of information on the predominant direction of biases as more taxa are added in, is certainly something to be aware of. It would be interesting to build the set of all "quartets" for a distance matrix, draw the split decomposition graph, then go back to the list of all quartets and remove just hose that are implied by the graph. Then build another graph showing as much as possible of the relationships amongst these remaining quartets and so on. A procedure like this would allow more detailed work on larger sets of taxa, without so much risk of missing important, but local biases.

3.9 ROBUSTNESS VIA CLASSIFICATION OF SITES INTO RATE CLASSES

In this section, the utility of classifying sites into additional rate classes is considered. This approach was inspired by the finding in table 3.4 that a mixture of two i.r. processes can give a substantial improvement in the fit between the observed and expected number of substitutions per site. Allowing two variable rate classes plus a proportion of invariant sites resulted in a G^2 statistic as good as any of the other models considered in table 3.4. Our assumptions, apart from the usually i.i.d. assumptions, are that sites that can undergo substitution follow one of two rate classes, and that the instantaneous rate matrix in one class is just a scalar of that in the other rate class. We also assume the number of sites in each rate is reliably estimated with methods like those used in table 3.4. Since the rate matrices in each rate class are multiples of each other, then conditional on a site having undergone x changes, we expect that the probabilities of the site patterns will be equal irrespective of the rate class the site belongs to.

For all sites showing x changes, construct a divergence matrix $\mathbf{F}_x$. Let the divergence matrix $\mathbf{F}_i$ for sites in rate class (i) be $\sum_x \mathbf{F}_x p(X = x / \lambda_i)$ be composed of dinucleotide pairs for sites showing x changes, weighted by the probability that a site showing w changes is from rate class i and not some other rate class. Under this model we conjecture the LogDet transformation

applied to any $\mathbf{F}_i$ will return distances which are additive in expectation (with both long sequences, and a sampling of sequences that allows an accurate gauge of each sites rate). It is then a mater of summing up the additive distances from each rate class to obtain an overall distance. In order to reduce variance we suggest that a tree selection procedure be applied to a $\Sigma w_i \mathbf{D}_i$, where $\mathbf{D}_i$ is a square matrix of pairwise distance estimates for the *i*-th rate class. An useful weighting, w_i, should be the inverse of the standard deviation of the sum of entries in $\mathbf{D}_i$, or some similar measure expressing the stochastic error in the distances for that rate class.

An important result helping to justify the above approach is, if two distance matrices are additive upon the same (unweighted tree), then any weighted sum of these same two distance matrices (**D**) will be additive upon a weighted form of the same tree (see proof in appendix 3.4). So if we can separate sites into their respective rate classes e.g. 1, 2, .., n (including sites with different underlying instantaneous rates) then as long as they evolved upon the same unweighted tree they $\mathbf{D}_1 + \mathbf{D}_2 + ... + \mathbf{D}_n$, can be added together prior to tree selection (as an aid to cutting down variance) and will be additive if each **D** matrix is additive, and hence allow accurate recovery of a weighted tree (where the weights may be interpreted as weighted averages of weighted averages of substitutions per site). This may be a useful way of adding together information from different molecules, if we expect they evolved according to the same unweighted tree, but at different relative rates in each lineage.

Exactly the same sort of procedure could be used to generate separate **s** vectors from each rate class, followed by the Hadamard conjugation, prior to adding together to give a single γ vector. A similar approach has been used by Van de Peer *et al.* (1993)(note however that they advocated using sites with deletions in as many as 25% of taxa, unaware of the non-additivity that such an inclusion can cause, nor did they consider the merits of weighting the distance matrices from each set of sites). These authors separated sites into 5 rate classes, using the observed pairwise similarities between disjoint groups of taxa to indicate the rate of each site. Barry and Hartigan (1987b) also considered a similar approach to making "asynchronous" distances more additive, by separating sites into first, second and third positions in proteins. The approach discussed here could be an improvement on classifying sites into very stringent rate classes, since we have more accurately estimated the set of sites which we expect would have one or the other intrinsic rates, by considering their expected parsimony length distribution. This type of approach is expected to become increasingly accurate as more sequences are aligned, since this is ought to give more information on the parsimony length of a site, and hence its expected relative rate. The observed number of changes on a tree (Farris 1969), likelihood (Olsen 1994), a tree independent compatibility weighting (Penny and Hendy 1985, 1986), or a pairwise distance comparison (Van de Peer *et al.* 1993) all offer useful ways of assigning sites into relative rate classes.

It is also possible to directly weight each site by its estimated rate, and then using these weights to form a weighted **F**(w) matrix, followed by applying the LogDet to **F**(w). This method is not guaranteed to be consistent. This is because while a weighted sum of transformations of observed distances for each rate class gives an overall additive distance, is not the same as the

log transformation of a sum of a weighted observed distances (analogous to the logarithm of a sum, not being the same as a sum of logarithms). However such a method could work well, especially since this type of weighting effectively ignores the information coming from more rapidly evolving sites in a graded way. It may be important not to give the slowest evolving sites too much weight however, since they could, due to the rarity of change, be marked by a high stochastic error in their frequency of occurrence. All these approaches should help to overcome the effects of unequal shifts of base composition in different rate classes, but those that accurately separate sites into distinct rate classes are expected to work best. Another type of weighting that might also yield useful approximate distances is analogous to generalised parsimony, and involves weighting of entries in **F** directly (e.g. scale down the size of the transition entries in **D** prior to forming **F**)(a suggestion by D. Penny). Such weighting should be carefully evaluated, and checked to see that it does not disrupt the robustness of the LogDet transform in real applications.

3.10 DISCUSSION

3.10.1 Earliest eukaryotic evolution reconsidered

This analysis suggests that the 16S-like rRNA data, when analysed under a more general model, provide support for Microsporidia being the sister group to all other living eukaryotes. With the removal of inferred invariant sites, the bootstrap support for this hypothesis is in the range of 80 to 92%, which may be considered strong evidence when the bootstrap is calibrated for type 1 and 2 errors (Rodrigo 1993) given an a priori hypothesis. It is important to consider other evidence that supports, or refutes, this conclusion. All Microsporidia, like *Giardia* (and other Diplomonadea) apparently lack mitochondria (and its possible derivative, the hydrogenosome), do not have well developed dictystomes on the Golgi body (involved in post translational protein modification) and lack peroxisomes (single membrane organelles which generate H_2O_2 for degradation of phenols and other cellular "waste") (e.g. see Cavalier-Smith 1993). These features suggest that these two lineages are more ancient than any other known eukaryotes, with the exception of possible relatives of the Diplomonadea, the Oxymonadea and the Retortamonadea, groups of flagellated protists that generally lack these features also. Attempts to culture these last two groups have failed thus far, and no sequences are reported from them as yet.

Further Microsporidia, like prokaryotes (and unlike *Giardia*) have 70S ribosomes, and the 5.8S rRNA sequence remains part of the 23S-like rRNA (Vossbrinck and Woese 1986). This last feature of a separate 5.8S rRNA, is unique amongst living eukaryotes, and since no prokaryote cleaves off the 5.8S rRNA fragment, this suggests it is a reliable phylogenetic marker indicating the branching of Microsporidia first from all other eukaryotes. Small 70S ribosomes are noted in only one other early eukaryotic group, this being the trichomonads, which appear to be more derived than Microsporidia and *Giardia* in having well developed Golgi bodies and peroxisomes. Trichomonads also have hydrogenosomes, which from sequence analysis appear to

be the result of an ancient symbiosis with a purple bacterium (Länge *et al.* 1994). Some suggest this organelle is a degenerate mitochondrion (e.g. Cavalier-Smith 1993), although direct evidence for this is lacking. Microsporidia also lack a suite of other morphological features even the diplomonads, trichomonads, oxymonads, and retortamonads have. Microsporidia lack lysozymes (involved in intracellular "digestion"), cilia and also centrioles (including those that form part of the spindle and centrosome, involved in meiosis and mitosis)(e.g. see Cavalier-Smith 1993). While it can be argued first two features may have been secondarily lost due to the extremely sessile intracellular parasitism that all Microsporidia practice, the third may well have not evolved amongst eukaryotes at the time Microsporidia diverged.

There is no good evidence to contradict a claim that Microsporidia are the most ancient diverging eukaryotes, suggestions discounting the above evidence seem to be based on the belief that standard sequence analyses do not lie (Cavalier-Smith's 1993 discussion seems to follow this line). It has been suggested that *Giardia* may have a Shine-Delgarno recognition sequence to initiate translation (e.g. Sogin *et al.* 1989). However, the leader sequence on mRNA's in *Giardia* are short, and so it appears unlikely prokaryotic processing is being used (M. Sogin pers. comm.). It remains unknown whether Microsporidia may harbour similar sequences, or could be using these actively, but this question deserves further study. As such there is no evidence here to suggest that *Giardia* and other Diplomonads, diverged earlier than Microsporidia. Some earlier studies of 16S-like rRNA sometimes put trichomonads deeper than Microsporidia (e.g. Leipe *et al.* 1993). However our own unpublished analyses of larger 16S-like rRNA data sets than those used here (results not shown) do not show trichomonads deeper than Microsporidia, a finding consistent with the earlier result being coloured by base composition effects; a possibility also acknowledged by the previous authors. Indeed there is some evidence from both sequence analyses (especially those placing emphasis upon transversional changes) and also morphology (Farmer 1993) that trichomonads (and their parent group the Parabasala) may form a monophyletic group along with diplomonads, retortamonads, and oxymonads.

Given what is apparently a strong prior probability on Microsporidia being the earliest diverging eukaryotes, it is surprising then that I have not seen this viewpoint championed in publication. Since Sogin *et al.*'s (1989) analysis which found strong bootstrap support for *Giardia* being deepest, it seems that biologists have only been to happy to take such findings as correct and explain away the contradictory morphological and biochemical evidence (Cavalier-Smith's 1993 does this). Worse still Cavalier-Smith (1993) put such faith in his analysis of sequences from the data base that he formally created and named many new phyla just because groups happened to appear on his tree. The analysis itself was inadequate, being an automated process with no serious attempt made to check the validity of the results (sequences were taken straight from data bases, put into an automatic alignment program with no use of secondary structure to verify alignments, all regions of the alignment were used, the only tree built was one based on Jukes-Cantor distances across all sites including large insertions and deletions, followed by neighbor joining). It is disappointing that phylogenetics and also systematics can be dominated by such superficial analyses, largely because the perception is that sequences always

contain much more information that other evidence. This misguided reverence to simple sequence analyses is repeated in the field of photosynthetic origins, where sequences experience similar strong shifts in base composition (see Lockhart *et al.* 1992) and also function (Lockhart *et al.* in press), yet are often taken at face value to overturn earlier arguments based on ultrastructure and biochemistry.

Despite these new findings we see no reason to reject the suggestion that the eukaryotes last common ancestors had an amoeboid form, although justification for this needs to be reestablished. Cavalier-Smith (1993) has supported this view, based largely on his assessment that vahlkampfiid amoebae would be the earliest branch in the eukaryotes. Sequences of these taxa by Hinkle and Sogin (1993), and our own analyses (not shown) make this appear unlikely. Nor does the hypothesis of "amoebae first" gain any direct support form trees such as that in figure 2.12. The problem of defining the form of the last common ancestor of living eukaryotes is partly that all described Microsporidia are so uniformly specialised in their morphology, it is difficult to make any guess as to their ancestral state. The diplomonads (plus retortamonads and oxymonads) and the trichomonads are also amitochondrial, but these are all flagellated forms. It is still unclear how these flagellates are related to one another, although as mentioned earlier, it is possible that they may form a single large monophyletic group. Farmer (1993) suggests this possibility based on specific ultrastructural features of some trichomonads and retortamonads. In nearly all trees after the divergence of these earliest exclusively amitochondrial groups are the unresolved relationships between amoebic forms (e.g. *Entamoeba*, *Physarum*, the vahlkampfids) and flagellated forms (e.g. the euglenozoa group including trypanosomes and *Crithidia*) with members having mitochondria, or at least evidence for mitochondria in their ancestors. My favoured interpretation is that the early eukaryote world was dominated by amitochondrial amoebic organisms that were the top predators on earth for a considerable period (at least several hundreds of millions of years) before (and probably after) increasing oxygen levels promoted endosymbiosis. One would expect that some such cells evolved to be quite large and possibly complex prior to the advent of mitochondrial forms, but that large amoebic forms (by then probably dominated by mitochondrial forms) demised with the rise of multicellular organisms. The Microsporidia may well have been a lineage that took to parasitising their larger cousins long before multicellular life evolved. Indeed some Microsporidia are known to parasitise single celled protists, including gregarines (e.g. Sprague 1977) which are possible relatives of the apicomplex group (a diverse group possibly just within the crown group, which include organisms such as *Plasmodium,* e.g. Sogin 1991).

The approximate time of origin of the Microsporidia can be made by noting that they branch approximately 2.5 times as deeply as the divergence amongst what is often called the crown group (the multicellular organisms and their close protist relatives). The crown group radiation from fossil data is typically put at about 1 to 1.4 billion years ago (e.g. Knoll 1994, Sogin 1991), suggesting that Microsporidia separated as long ago as 2.5 to 3.5 billion years. There have even been claims of multicellular eukaryotes as old as 2.1 billion years (Han and Runnegar 1992), and if substantiated then the implication must be that Microsporidia separated from other eukaryotes

very early in the history of the earth.. This, of course, assumes approximately equal rates of evolution amongst the eukaryotic taxa used to make this assessment. This assumption of clock-like behaviour is compromised when the substitution process varies significantly as in the case of the earliest eukaryotes studied (exemplified by varying GC contents). Further the earliest eukaryotes have also probably evolved faster than the latter eukaryotes. While the LogDet distances probably have a bias towards overestimating the relative external edge lengths to *Giardia* and *Vairimorpha*, their parasitic life styles would be consistent with higher substitution rates as a selective factor towards avoiding host defenses (this will become more clear when as more representatives of these groups are sequenced). However even with these possible violations, when the tree is rooted outside the eukaryotes (which is almost certainly the case), then the similar edge lengths suggest that to a first order of approximation this extrapolation is probably still valid. Further a clock-like calibration could be quite valid if the "back-bone lineage" was not experiencing major base composition changes. Determining the accuracy of this assumption would require further sequences and a more accurate inference of ancestral base compositions of this molecule during early eukaryotic evolution.

Microsporidia being deepest amongst the eukaryotes has interesting implications for the origins of sex. One category of Microsporidia have alternating paired and unpaired nuclei in different stages of their life cycle. Meiosis is known only from these species, which have a complex sexual cycle (e.g. Canning 1994). Given the complexity of this process and the fact that a few early groups including the diplomonads and some of the early amoebae, show no evidence of meiosis, makes it appear probable the Microsporidia separately evolved sexuality. This in turn appears to suggest that the last common ancestor of living eukaryotes if sexual at all, was not using the cellular structures seen in the more familiar eukaryotes. This may explain why Microsporidia lack centrioles, which are part of the meiotic apparatus of other eukaryotes.

3.10.2 The need to study variances and bias

It is important that we do not become complacent with the methods we have, but should seek to understand their performance in detail. One area that deserves further study are the sampling and systematic errors that occur with distance transformations. In this chapter we have already illustrated the very strong systematic errors that can occur in either direction (over or underestimation) when standard stationary model distance transformations are applied to data with non-stationary base compositions. We also have discussed how the LogDet will also show systematic error, particularly an inability to fully account for base composition shifts when estimating distances between sequences where sites evolve at different rates. More study should also be done on the sampling errors of distance estimates, with a view to improving their performance if possible (e.g. Schöniger and von Haeseler 1993).

In recent unpublished studies with P.O. Lewis and D.L. Swofford it was observed that the sampling error of the LogDet transform generally stops having an significant effect on tree selection as soon as the sequence length becomes reasonable (for example > 200 sites). By looking at trees with long edges in various configurations (see chapter 5), it is possible show that

sampling bias is also dropping away quickly in its effect upon tree selection. It will be important to study if systematic or sampling bias is responsible for certain features of tree selection, for example the very low bootstrap support for Microsporidia and *Giardia* together which have very different base compositions (although this could also be due to a bias in the neighbor joining algorithm).

To look at ways of improving the sampling properties of the LogDet, which already seems very favourable, it is important to study the various components of the total error. The sampling bias in the LogDet is a combination of two main factors, the bias in the estimation of the determinant, and the bias that is imparted by the non-linear log transform. It is interesting to study the contribution of each, as well as a possible third factor, which is due to the base frequency correction terms (only in equation 3.3.1-2). Preliminary results (simulations with D.L. Swofford and P.O. Lewis) show that the overall bias, which is generally towards overestimating the true distance, drops away quickly as c increases, although not as quickly as with the simpler standard distances, e.g. the Kimura 3ST formula (when compared under that model and submodels). Simulations by D. Penny (pers. comm.) suggest that a bias towards underestimating the determinant falls away even more quickly with increasing c. Logically then, this leaves a higher sampling variance to the determinant (verses the eigen values of the Kimura 3ST equation, e.g. [1-2P-2Q]), combined with the nonlinear logarithmic transform, as the reason for the bias persisting longer than with the other methods. A fourth source of bias will come into play when distances are larger than expected under the model, and will be dependent upon how such large distances are treated. We have already shown (see section 3.4.7) that by altering the distance value that is given to negative determinants in bootstrap samples, that this source of bias does not appear to have much effect upon tree selection (other studies in chapter 4 consider this bias with other methods). Studies can also be carried out on ways of estimating the variance of the LogDet, with Bar-Hen and Penny (1996) looking at bias in variance estimates generated by the jackknife and the bootstrap. However, studies in Waddell *et al.* 1994 (and chapter 4) suggest that when working with a sample the delta method approximations (such as equation 3.5.1-1) may offer both convenient and minimum bias estimation of variances, so it is important to study these estimators (e.g. equation 3.5.1-1) also.

3.10.3 Miscellaneous discussion

Our aim has been to better control for what are often considered the three major causes of inconsistency in tree building; unequal edge lengths in the tree, unequal rates at different sites, and an unknown form of substitution matrix (and especially a non-stationary underlying process of substitution leading to nucleotide composition differences). After doing this much of the tree remained stable, often giving fairly strong resolution of internal edges. However the more closely we study the actual molecules, and either different subsets of sites or taxa, the more it becomes apparent that there remain important non-i.i.d. effects in the evolution of these molecules which are capable of leading to inconsistency. In the case of early eukaryotic divergences (and almost certainly the case amongst the bacteria also) much more care and fundamental research needs to go into constructing reliable phylogenies based upon both

sequences and morphology. Clear examples of phylogenetic sequence analysis being woefully inadequate with regard to conclusions are evident in Cavalier-Smith (1993) and Sogin *et al.* (1989). Kabnick and Peattie (1991) then built further conclusions upon the later paper. The analyses by Siddall *et al.* (1992) are a good example of dismissing some of the dubious analyses that have been published in this area recently. Much useful work remains to be done even with the present sets of sequences (e.g. rRNA's, elongation factors, RNA polymerase, α and β tubulins) to explain why even these few data sets do not seem to be agreeing on early eukaryote evolution within expected sampling errors.

An important general implication of the work in this chapter is that there appear to be at least 160 sites (or at least 20% of this data set) in 16S-like rRNA sites which should not be used in any phylogenetic analysis of major taxonomic groups. This conservative "20%" estimate being made by the mixed invariant sites-Γ model in table 3.4. It has yet to be studied whether this statement holds also for the 16S-like sequences of mitochondria and chloroplasts, or when comparing them to nuclear encoded genes. Given that 16S-like rRNA is evolving by some sort of covarion process, then some of the invariant sites in the nuclear copy of this gene may be changing at a perceptible rate amongst the highly divergent organelle rRNA's. It could turn out to be problematic for mixed analyses of organelle and nuclear 16S-like sequences (which are used especially in deriving the origins of chloroplasts and mitochondria) if the invariant sites each type did not match up.

Identifying, or even counting, the number of these invariant sites in genes with many available sequences could be greatly compromised by sequencing errors. Assuming the rate of such errors to be as low as 1 in 1000, then across all the sequences in the current databases (over 4,000 for 16S-like rRNA) there would be approximately 4 errors per site. While good sequencing studies have been shown in retrospect to have error rates as low as 1 in 3000, some recently published 16S-like rRNA sequences from Microsporidia have error rates as high as 5%, or 50 per 1000, including missing bases! (N. Pieniazek, pers. comm., based on resequencing others studies). Hopefully the use of phylogenetic information and secondary structures can weed out many of these sequencing errors in slowly evolving sites, which, if errors are random should not occur in closely related species. It must be useful to phylogeneticists to have these sites identified so that they can readily be removed from an analysis, say, of arthropod origins, prior even to any further corrections for unequal rates between sites. Their removal will not only help on this account but also allow researchers and methods including LogDet transforms and maximum likelihood to "see" base compositional biases more easily.

The identification of the problems caused by different rate classes (including constant sites) having different base compositions, that will shift relative to one another through time is important. All methods of phylogenetic analysis (including maximum likelihood) must attempt to take this into account when it becomes significant. A second effect which may also need to be considered is unequal base compositions between regions of a molecule or between different molecules. Vawter and Brown (1993) for example have shown that stem and loop regions can have a quite distinct base composition from bulges, and loops. While it may be splitting the data

up to much to do this when analysing just 16S-like rRNA sequences, it should be considered in larger data sets. For example adding together 16S-like, 23S-like and 5S-like rRNA sequences will often give over 4000 base pairs (although dropping to less than 2,000 for the deepest divergences) and we could consider making these separate classes even when using invariant site-LogDet transforms as each classification could then have over 500 sites, even with additional separation of sites into a few rate classes.

Some my ask, "was it really necessary to reduce the data set down to so few sites by removing regions containing insertions and deletions?" The answer is yes, given the reliance upon i.i.d. models when making inferences for very anciently diverged molecules, especially studies which require rooting from another group (e.g. rooting the eukaryotic tree with archaebacteria). A major concern with some recent analyses are that they include badly aligned regions. In addition even when a correctly aligned insertion or deletion region is used to help estimate the distance to some but not other groups, this is almost certainly introducing extra non-additivity (it is very much like comparing "apples and oranges"). In addition, at least for the deep divergences in the tree, insertion / deletion regions frequently appear to be evolving more rapidly than the average of the remaining variable sites. The main resistance to removing these sites are (1) it makes my sequences look short, (2) it may reduce the bootstrap support for my group of interest, and (3) it is throwing away data (albeit data which is expected to distort resolving the deepest parts of the tree). Of these three sources of resistance, only the third is a valid objection, and how it impinges upon 2.

To overcome objection (3) we suggest a compromised form of data editing and analysis which retains the most reliable information in a distance analysis. From the set of all sequence sites, build trees for subsets of the taxa which do not have a major insertion / deletion problem. From these analyses determine which parts of the tree you are very confident in. If it turns out that insertions or deletions are supporting these edges, or quite unambiguously supporting other edges, use this information to specify a set of constraints on the optimal tree when using all the edited sequences (programs such as PAUP*, Swofford 1995 allow such constraints and include distance and ML analyses). Next remove all the poorly aligned and indel regions, plus any regions which are accumulating too many changes for estimating the deepest divergences, and run the tree selection with the constraints in place. By doing this you lost a minimal amount of information and at the same time made the sensible move to make distances additive.

Clearly it is not reasonable to expect standard distance transformations, or the LogDet, to restore additivity if some of your sequences have less than 50% of sites in common and some of these of dubious alignment. If it is not certain that an insertion or deletion event is a reliable marker, then the difficult decision of whether to specify a constraint must be made. This is of course preferable to throwing in sites affected by insertions and deletions and hoping for the best. Even if a potentially informative insertion or deletion is not used to specify a constraint on the optimal tree, a search of the partitions in the bootstrapped trees is reasonable. This could indicate if it was likely to have been defining a correct group, and this can be mentioned in the

results. In future an integrated and realistic likelihood approach for including such information may be possible (see Kishino *et al.* 1990 for an attempt to do just this).

3.10.4 Base composition bias and biased substitution

It is important to distinguish two features which could be grouped under the heading of "base composition problems". One is the overall base composition of a sequence, which causes the gross effect of grouping of sequences by base composition (e.g. Lockhart *et al.* 1992), or the potential strong repulsion of sequences when standard distance transformations are applied (section 3.4.3). To a large extent the invariant sites-LogDet transform should counter this sort of effect, although with unequal shifting of base composition for sites in different rate classes, it may be necessary to separate sites into more rate classes prior to making separate transformations. A second effect which is much harder to detect and presently little understood is a more specific type of biased substitution. If sequences are generally very slowly evolving and have many near invariant sites, it may require relatively few parallelisms or convergences at the variable sites to mislead all current methods. One possible reason for a high proportion of misleading changes would be substitutions in a small subset of sites, coupled with different taxa having strong and distinctive substitution preferences. *Giardia* and *Vairimorpha* show evidence for strong and distinct substitution, which is accompanied by gross base composition shifts. However *Physarum* hints at biased substitution at the singleton sites, yet overall has a base composition frequency very like that of most other eukaryotes.

An important challenge is to understand how such biased substitution, coupled with covarion evolution (Fitch and Markowitz 1970) which could open up suites of previously invariant sites to substitution under particular conditions, can be detected and diagnosed. The cause of the distinct attraction between the nematode and other more ancient eukaryotes (particularly *Crithidia* and *Vairimorpha*) is not yet apparent, but something of this nature could be the explanation. The possibility that a combination of shifting base composition, rates across sites, and covarion types of evolution could effectively hide very strong substitution biases, needs further investigation.

The evidence from the perfect sites evaluations (section 4.8.1) suggest that some sites which were only evolving during the early part of the eukaryotic tree, showed strong base composition bias (leaving just the transversional changes as clearly informative) and but did not undergo transversional changes again over the rest of the "tree of life". It is easy to imagine that similar processes could lead to apparently slowly changing sites being subject to misleading changes in the critical period when they were recording evolutionary history. With shifting variable sites and shifting substitution biases, it is possible that such locally strong biases around certain internal edges could go largely "unnoticed" and unaccounted for by all current methods, including the LogDet. This problem could be worse for coding sequences if they undergo preferred changes (for example in the same amino acid class), with what might also be strong codon preferences. Thus intermittently variable codons might well have just two possible states when they are variable, making parallelisms and convergences relatively common place.

While we presently look for similar base compositions leading to sequences grouping together, it is also important to consider what happens when two near relatives begin to experience nucleotide content changes in opposite directions. If the internal edge linking them is not long it is easy to see that standard distance methods in particular (but other methods also) could decisively overestimate the distance between them (relative to the estimate of distances between stationary sequences, see section 3.4.3). Consequently we have an "anti-Felsenstein zone" scenario due to edges repelling each other (see chapter 5), and a rapidly climbing probability of separating these two sequences the longer this pattern of evolution occurs. This possibility must be considered in real sequences. It may, for example, be part of the reason why distance methods very rarely group *Giardia* and *Vairimorpha* even by chance in 1000's of bootstrap replicates. Overall it is fair to say we are only beginning to understand how convergences and parallelisms may perturb tree estimation, and as yet we have only vague ideas of how to diagnose what is occurring in real sequences which do not follow simple i.i.d. models. Accordingly, it is unwise to be over-confident in ruling out problems just because the more sophisticated i.i.d. tree inference techniques (including LogDet or ML) give similar answers to earlier methods.

3.10.5 Invariant sites-LogDet transforms: A most useful distance estimate

Overall then, invariant sites-LogDet transformations would seem to offer distances that are more additive over a much wider range of models than the standard distance transformations. It is also very pleasing to find that the sampling variance of these transformations does not appear to be any more serious a concern than using standard distances, as long as sequences are at least some hundreds of bases long. Indeed, in our examples the invariant sites-LogDet transform was giving trees with very similar overall bootstrap support to the simple maximum likelihood Poisson distance estimator (the Jukes-Cantor distance estimate). In contrast, a real concern is that the application of standard stationary distance estimators can make the situation markedly worse than using observed distances when base composition is non-stationary. If one is going to use these estimators, then it would seem desirable to make "infinite distances" modifications to them (see section 3.3.2) (this applies double to distances allowing for a distribution of rates across sites, e.g. those of Golding 1983, Olsen 1987, Jin and Nei 1990). It will be important to study the performance of these estimators in simulations, as well as compare them with invariant sites-LogDet transforms on real studies. If stationarity of sequences can be established, then with long sequences it may well be desirable to use the general time reversible distance with an appropriate distribution of rates across sites (equation 3.2.2-1). The issue of different rate classes having different base compositions is very important, and if phylogenetic methods are to claim to be robust to both base composition and unequal rates across sites they must consider this factor (this goes also for maximum likelihood methods, which should not automatically be assumed more robust than other methods).

Appendix 3.1 Proof that all 2-state transition matrices can be considered the result of a continuous time process

We show that every 2-state probability transition matrix with positive determinant can be expressed as the result of a continuous time process. We define **P** as,

$\begin{bmatrix} 1-a & a \\ b & 1-b \end{bmatrix}$, where a and b are each zero or greater, but less than 1.

For a continuous time Markov process, $\mathbf{P} = \exp(\mathbf{R}t)$, which rearranges to $\ln(\mathbf{P}) = \mathbf{R}t$, where **P** is a matrix of all positive values such that each row sums to 1, **R** is an instantaneous rate matrix (off diagonal elements ≥ 0, and rows sum to 0), t is a scalar (not necessarily linear with time), and exp is the matrix exponent, while ln is its inverse function. The matrix log of **P** will be interpretable as the result of a continuous time process if $\ln(\mathbf{P}) = \Omega\ln(\Psi)\Omega^{-1}$, where Ω is a matrix containing as columns the eigenvectors of **P**, Ω^{-1} is its inverse, while Ψ is a diagonal matrix of the eigenvalues of **P**, and the logarithmic function is applied to the diagonal entries of componentwise. This condition holds if **P** has two distinct and positive eigenvalues. The characteristic equation of **P** is, $0 = (1\text{-}a\text{-}x)(1\text{-}b\text{-}x)\text{-}ab = (1\text{-}a\text{-}b) + (-2+a+b)x + x^2$, which is obtained by subtracting x from each diagonal element, then setting the determinant equation to zero. Since the roots of x are the eigenvalues of **P**, it remains to be shown that they are always distinct and positive. Recalling that $\det(\mathbf{P}) > 0$, so $(1\text{-}a)(1\text{-}b)\text{-}ab > 0$, thus $1\text{-}a\text{-}b > 0$, while $(-2+a+b) = -1\text{-}(1\text{-}a\text{-}b)$. Consequently $(1\text{-}a\text{-}b) + (-2+a+b)x + x^2$ always factorises as $((1\text{-}a\text{-}b)\text{-}x)\ (1\text{-}x)$, given our constraint that $1\text{-}a\text{-}b>0$, and a and $b \geq 0$. That the roots are distinct as long as there is any change follows because the discriminant $= (-2 + a + b)^2 - 4(1 - a - b) = (a + b) > 0$ (provided a or b > 0). Thus the characteristic equation always has two positve roots, for **P** with determinant > 0, completing the proof.

Appendix 3.2 Proof of the identity of averaged "asynchronous distances," LogDet and paralinear distances

Barry and Hartigan symmetrised their asynchronous distances when doing a practical example, but gave no justification for this, having based their proof rather on the consistency of $-1/4\ \ln[\det(\mathbf{P}_{ij})]$ or $-1/4\ \ln[\det(\mathbf{P}_{ji})]$. Recalling that $\mathbf{P}_{ij} = \pi_i^{-1}\mathbf{F}_{ij}$, and $\mathbf{P}_{ji} = \pi_j^{-1}\ \mathbf{F}_{ij}^t$ (Barry and Hartigan 1987, and where t defines transpose), then,

$$1/2(-1/4\ \ln[\det(\mathbf{P}_{ij})] + -1/4\ \ln[\det(\mathbf{P}_{ji})]) = -1/8(\ln[\det(\mathbf{P}_{ij})] + \ln[\det(\mathbf{P}_{ji})])$$

$$= -1/8(\ln[\det(\pi_i^{-1}\mathbf{F}_{ij})] + \ln[\det(\pi_j^{-1}\mathbf{F}_{ij}^t)])$$

(matrix determinants are multiplicative)

$$= -1/8(\ln[\det(\pi_i^{-1})\det(\mathbf{F}_{ij})]+\ln[\det(\pi_j^{-1})\det(\mathbf{F}_{ij}^t)]),$$

(determinants are commutative)

$$= -1/8(\ln[\det(\mathbf{F}_{ij})\det(\pi_i^{-1})] + \ln[\det(\mathbf{F}_{ij}^t)\det(\pi_j^{-1})])$$

(since $\det(\mathbf{M}^{-1}) = 1/\det(\mathbf{M})$)

$$= -1/8(\ln[\det(\mathbf{F}_{ij})/\det(\pi_i)] + \ln[\det(\mathbf{F}_{ij}^t)/\det(\pi_j)])$$

(taking logs of terms in [ ] brackets)

$$= -1/8(\ln[\det(\mathbf{F}_{ij})] -\ln[\det(\pi_i)] +\ln[\det(\mathbf{F}_{ij}^t)] -\ln[\det(\pi_j)])$$

(transposing doesn't change determinant) $= -1/8(2\ln[\det(\mathbf{F}_{ij})] - \ln[\det(\pi_i)] - \ln[\det(\pi_j)])$

[a convenient form for computation] $= 1/4(-\ln[\det(\mathbf{F}_{ij})] + 0.5\ln[\det(\pi_i)] + 0.5\ln[\det(\pi_j)])$

$= 1/4(-\ln[\det(\mathbf{F}_{ij})] + \ln[\det(\pi_i\pi_j)] / 2)$

(taking minus sign outside) $= -1/4\,(\ln[\det(\mathbf{F}_{ij})] - \ln[\det(\pi_i\pi_j)] / 2)$

(taking logarithm of inner expression) $= -1/4\,\ln(-[\det(\mathbf{F}_{ij})] / [\det(\pi_i\pi_j)]^{1/2})$

$= -1/4\,\ln([\det(\mathbf{F}_{ij})] / [\det(\pi_i)^{1/2}\det(\pi_j)^{1/2}])$.

Notice that the penultimate form is the same as equation (3) of Lockhart *et al.* 1994 (the standardised form of Steel's (1994) $-\ln[\det(\mathbf{F}_{ij})]$). The final arrangement is the same as that of Lake (1994) but for his neglecting the factor of 1/4. The form fifth from last is probably the most convenient computationally (in terms of operations used, and avoiding round off error).

Appendix 3.3 Proof that F is symmetric under time reversibility and the clock

Proof that under a time reversible model, with the root distribution stationary, **F** is always symmetric. (N.B. here $\mathbf{A}^t$ denotes transpose of **A**).

If **R** is reversible, $\mathbf{F} = \exp(\mathbf{R}t_1)^t \Pi \exp(\mathbf{R}t_2)$

$= \Pi\exp(\mathbf{R}(t_1 + t_2)$

$\mathbf{F}^t = \exp(\mathbf{R}t_2)^t \Pi^t \exp(\mathbf{R}t_1)$

$= \exp(\mathbf{R}t_2)^t \Pi \exp(\mathbf{R}t_1)$

$= \Pi \exp(\mathbf{R}t_2)\exp(\mathbf{R}t_1)$

$= \Pi \exp(\mathbf{R}(t_1 + t_2))$

so, $\mathbf{F} = \mathbf{F}^t$.

Proof that for any Markov process operating on a rooted tree obeying a molecular clock, then F is symmetric (no assumption about root being in equilibrium).

$$\mathbf{F} = \mathbf{P}^t\Pi\mathbf{P}$$

$$\mathbf{F}^t = \mathbf{P}^t\Pi(\mathbf{P}^t)^t = \mathbf{F}.$$

Since **F** has at most 9 degrees of freedom under a molecular clock, yet **R** can have up to 12, it follows that we cannot solve for **R** given just **F** under the most general i.i.d. model. Hence we cannot solve for $\delta = -\mathrm{tr}(\Pi\mathbf{R})$ (the number of substitutions per site. However we can solve for trace of **R** using the LogDet equation of Steel *et al.* (1993) or Lake (1994), and as long as the Π matrix of the root distribution is in equilibrium with **R**, then this measure will be collinear with time. Hence this LogDet distance is quite suitable for divergence time calibrations on any tree reconstructed from data obeying this model. Analogous to the case with the time reversible model, if LogDet is to be used for reconstructing trees under a molecular clock, then **F** should be symmetrised to reduce its (and the final distances) variance. These results are due to

collaborative research with Dr Mike Steel, and are contained in Waddell and Steel (in preparation).

Appendix 3.4 Proof that any two distance matrices additive on the same unweighted tree are still additive when linearly combined

Proof that any two distance matrices additive upon the same unweighted tree (but possibly with quite unlike edge weights), when summed are also additive upon the same unweighted tree.

Let $\mathbf{D}_1 \times x$ be an additive set of distances upon tree T_1, and $\mathbf{D}_2 \times y$ be additive upon T_2 (and x and y are both ≥ 0). Now T_1 and T_2 are the same unweighted tree if they have exactly the same set of edge indices. So let $\mathbf{w}_1$ be the ordered vector of indexed weighted edges in T_1 and let $\mathbf{w}_2$ be the ordered, indexed set of edge weights in T_2. Let $\mathbf{w}_1 + \mathbf{w}_2 = \mathbf{w}_3$; clearly by the additive property of vectors, $\mathbf{w}_3$ is the set of weighted edges defining tree T_3. Thus $x\mathbf{D}_1 + y\mathbf{D}_2$ is additive upon T_3, and T_3 by having the same set of edge indices as T_1 and T_2, is the same unweighted tree, completing the proof. (Note: Proof can be extended so x and / or y may be negative, but this makes little sense in the current context). (Thanks to Jeff Thorne for inciting this proof).

CHAPTER 4:

SAMPLING ERRORS ASSOCIATED WITH TRANSFORMED DATA

4.1 INTRODUCTION

A major priority in phylogenetic research is to assess the statistical reliability of trees built from sequence data (Felsenstein 1988; Penny *et al.* 1992). For trees estimated from distances, the variances and covariances of edge lengths are important in quantifying the sampling errors arising from finite sequence length (e.g. Hasegawa *et al.* 1985, Nei and Jin 1989, while Bulmer 1991a provides a useful summary of standard least squares techniques on trees). Here we describe equivalent results for sequence data corrected for multiple changes by Hadamard conjugations (chapter 2). With this method, aligned sequences are corrected for implied multiple changes (hits) prior to choosing an optimal tree(s).

The proportions of sites which show each pattern of character states are expressed as a vector $\hat{\mathbf{s}}$ ($\mathbf{s}$ from sequences, ^ being a sample estimate of pattern probabilities). If sites are independent, then the patterns observed in a sequence of length c are a random sample from a multinomial distribution (Felsenstein, 1981; 1982), with parameters c and $\mathbf{s}(T)$ (the vector of model probabilities of observing each pattern).(Thus $\mathbf{s}(T)$ may be considered the population mean vector, while $\hat{\mathbf{s}}$ is a sample). Assuming a sample form a multinomial distribution, it is possible to estimate $\mathbf{V}[\hat{\mathbf{s}}]$, the variance/covariance matrix of $\hat{\mathbf{s}}$. The sample error in $\hat{\mathbf{s}}$ is transformed by the Hadamard conjugation into error in $\hat{\gamma}$. In this chapter we show how to calculate the variance-covariance matrix $\mathbf{V}[\hat{\gamma}]$ for the entries in $\hat{\gamma}$, and from this derive the size of the correlations of pairs of entries in $\hat{\gamma}$.

The variances and covariances of $\hat{\gamma}$ give an approximate measure of the sampling error of an entry in $\hat{\gamma}$. To know precisely the probability of an error, we need to know the form of the marginal distributions of $\hat{\gamma}$ (the values $\hat{\gamma}_i$ takes when $\hat{\mathbf{s}}$ is a random sample from $\mathbf{s}(T)$). For example, are the marginal distributions Poisson ? Normal ? etc. The bivariate distributions of entries in $\hat{\gamma}$ will show if specific pairs are linearly correlated. It is already known that $\hat{\gamma}$ is a consistent estimator (Hendy and Penny 1993, Steel *et al.* 1992 chapter 2) under a variety of identical rate (i.r.) models. That is $\hat{\gamma}$, estimated from longer and longer sequences, will converge to $\gamma(T)$ a vector representation of the original weighted tree, T. Since the Hadamard conjugation, like distance corrections, is a non-linear transformation we expect a degree of statistical bias (a systematic departure of the mean of $\hat{\gamma}$ from $\gamma(T)$) when applying it to short sequences (Stuart and Ord 1987, p. 324). Using simulations we estimate the bias in $\hat{\gamma}$ when it is calculated with sequence lengths commonly used in phylogenetic studies.

Two factors effecting the form of the multivariate distribution of $\hat{\gamma}$ are c (the sequence length) and the rate of change per site (λ). Parameter c will influence the magnitude of the error in $\hat{\mathbf{s}}$ (relative to entries in $\mathbf{s}(T)$), while λ will determine how many multiple changes are expected. We infer the effects of c and λ on the distribution of $\hat{\gamma}$ by considering the intermediate steps in the Hadamard conjugation from $\hat{\mathbf{s}}$ to $\hat{\gamma}$. The changes in the multivariate distribution of $\hat{\gamma}$ are illustrated with simulations at selected values for c and λ. Note that in this chapter, λ is a label referring to the intrinsic rate of all sites, which are also assumed i.r in section 4.2 to 4.4. For these sections, factor λ only changes when we consider another scaled version of the same weighted tree (i.e. the same weighted tree times a scalar e.g. see figure 4.1).

The delta method (Stuart and Ord 1987, p. 324, Kimura and Ohta 1972, Bulmer 1991a, Rzhetsky and Nei 1992), a convenient first order approximation used to estimate the variances and covariances of variables after a non-linear transformation, is also expected to cause bias in estimating $\mathbf{V}[\hat{\gamma}]$ (Stuart and Ord 1987, p. 324). We examine the magnitude and pattern of this bias when calculating population and sample variance-covariance matrices of $\hat{\gamma}$.

This chapter is large and contains a number of distinct, but related, themes. The first four sections of this chapter deal with understanding the sampling structure of the i.r. Hadamard conjugations. This work has been published as Waddell *et al.* (1994) "*The sampling distributions and covariance matrix of phylogenetic spectra.*" The next section, 4.5, extends these results to include unequal rates across sites (URAS) correction formulae. It reveals some surprising results, which are relevant to all methods which explicitly allow for URAS effects, including distance transformations and ML. It is shown that when considering the issue of weighting transformed data sets prior to their recombination, there comes a definite point where it is best to remove rapidly evolving sites, rather than try to incorporate them in the model or give them low weight. This is shown to be the case using sampling error alone, but is reinforced by the expectation of a positive correlation between systematic errors and the amount of evolutionary change sites undergo. Similar findings are expected for predictive-fit methods, including ML tree selection.

Section 4.5 also indicates that given long sequences, surprisingly large distances ($\approx$ 2 substitutions per site), can be more informative than moderately short distances (taking into account random, but not systematic errors). Retaining sites showing many changes goes against the traditional wisdom of excluding sites such as third positions (e.g. Horai *et al.* 1992) if they are suspected of having an average of, say, 1 change per site. In contrast, Waddell and Penny (1995)(see also chapter 5) retained the third position sites in their 4-state ML analyses of the Horai *et al.* (1992) data (because there was still some clear "tree signal" amongst them e.g. see figure 2.8). Conversely, it would probably be unwise to retain all mtDNA third position sites when making 4-state ML analyses of the origin of mammals. Clearly, an important question which has yet to be seriously addressed are procedures to include / exclude sites for a specific analysis. While general guides for classes of sites are fairly straightforward (e.g. do the third position sites well differentiate the tree), it is harder, for example, to know when specific first or

second position sites should be excluded because they are evolving too fast (or conversely when only slowly evolving third position sites should be retained). The studies in this section are step towards this goal.

Section 4.6 considers ways to reduced the sampling variance of Hadamard conjugations, which tend to be higher than those for just pairwise distance transformations. These include a way of forcing the 4-state Hadamard conjugation to the generalised Kimura 2ST model, or the Jukes-Cantor model. Important extensions of these restrictions are the identification of families of linear tree invariants under the Kimura 2ST and Jukes-Cantor models, which exist for any number of taxa. Another exciting finding is that reduced bias estimators of the form introduced by Tajima (1993a) are applicable to Hadamard conjugations. Further study of these, and other novel estimators in appendix 4.2, show that they may offer substantial reductions in sampling variance. This, rather than their unbiased nature, is expected to be the real advantage to their use. An additional result is that the delta method of estimating the variance of these estimators can be quite inaccurate with very short sequence lengths ($c < 100$).

Section 4.7 considers sampling properties of the distance Hadamard, a method for going from just pairwise distances (δ) to an analogue of γ. This method, introduced by Hendy and Penny (1993), estimates pathset lengths indirectly as a minimal sum of pairwise distances with the same endpoints. It is an intriguing method, being able to use any pairwise distance data, yet little is known of its statistical properties. The distance Hadamard is known to be consistent (in estimating a vector description of the true tree) if the pairwise distance estimates are additive upon the true tree (Hendy and Penny 1993). It is called the "distance Hadamard transform", or just "distance Hadamard," (rather than a "Hadamard conjugation") because only one Hadamard transform is used. The distance Hadamard starts with a pairwise distance matrix (usually already "corrected" for multiple substitutions), estimates all the pathset lengths between more than two end points using just pairwise distances (to generate the $\rho(d)$ vector), then applies the inverse Hadamard transform ($\mathbf{H}^{-1}$) to estimate $\gamma(d)$. The variance covariance matrix of $\hat{\gamma}(d)$ is derived, and using simulations, the sampling distribution of $\hat{\gamma}(d)$ is also evaluated. Overall the vector $\hat{\gamma}(d)$ is found to have reduced sampling variance relative to $\hat{\gamma}$, and the reasons for this are explained. Later, in chapter 5, the results in this chapter are related to the statistical efficiency of tree selection from $\hat{\gamma}$ verses tree selection from $\hat{\gamma}(d)$.

The final results section, 4.8, considers the perplexing question: What sort of ML estimators, if any, are $\hat{\gamma}$ vectors? This section concludes by relating the sampling properties of $\hat{\gamma}$ to the statistical efficiency of tree selection.

4.2 CALCULATING THE VARIANCE-COVARIANCE MATRIX OF PHYLOGENETIC SPECTRA

4.2.1 Our illustrative model

For the next few sections (4.2.1 to 4.3), we will use the following four taxon example to generate a starting **s** vector and to illustrate our calculations. Following Hendy and Penny (1993) let our weighted tree, T of t taxa, be described by the vector of weighted edge lengths $\gamma(T)$. The tree T we use to illustrate our methods (fig. 2.1, chapter 2) is similar to Felsenstein's (1978a) example with only two distinct edge weights which are chosen so that parsimony applied to uncorrected sequence data will converge to the wrong tree. The edge lengths (weights) are the total expected number of changes per nucleotide site (counting multiple substitutions). These are represented by $\gamma(T)$ of figure 2.1b The mechanism of character state change used here is that of Cavender (1978) and Hendy and Penny (1989), the 2-state analogue of the Jukes-Cantor equation. Changes at sites occur independently of each other and the probability of changing is identical at all sites, so the mechanism of change is i.i.d. (independent and identically distributed) and i.r. (identical rates at each site). The use of λ in the first sections (4.2 to 4.4) of this chapter means the intrinsic rate with respect to total amounts of change (not relative rates among sites). We assign $\lambda = 1$ (or λ_1) to the sites on the weighted tree in figure 2.1. If we scale this tree by 1/2, then we assign $\lambda_{1/2}$ to sites evolving on this tree etc.

Following Hendy and Penny (1993) we apply the Hadamard conjugation to $\gamma(T)$ to obtain **s**(T) the vector of the probabilities for each of the possible 2^{t-1} site patterns (bipartitions), that is

$$\mathbf{s}(T) = \mathbf{H}^{-1}(\exp(\mathbf{H}\gamma(T))). \qquad (4.2.1\text{-}1)$$

In this formula **H** is a symmetric Hadamard matrix of 2^{t-1} rows, $\mathbf{H}^{-1} = 2^{-(t-1)}\mathbf{H}$, and the exponent function is applied component-wise. The calculations of equation (4.2.1-1) are fully invertible. Recalling that $\gamma(T) = \mathbf{H}^{-1}(\ln(\mathbf{H}\mathbf{s}(T)))$, the calculations of the variance-covariance matrix of $\hat{\gamma}$ are broken up into a series of intermediate steps and give the results of a numerical example, starting with **s**(T) and its covariance matrix in table 4.1.

The vector **s**(T) (shown in table 4.1, and as the grey columns or $\lambda = 1$ in figure 4.1b) is calculated from the $\gamma(T)$ of figure 2.1b using equation 2.2.1-1. The frequencies of the bipartitions in a sequence of length c are, under this model, a random observation from a multinomial distribution (Felsenstein 1981a) with parameters c and **s**(T). Vector $\hat{\mathbf{s}}$ is vector $\hat{\mathbf{f}} / c$ (where $\hat{f}_i$ is the observed frequency of the i-th sequence bipartition) and so is our maximum likelihood estimate of **s**(T). Consequently $\hat{\mathbf{s}}$ has the distribution of the above multinomial scaled by 1/c.

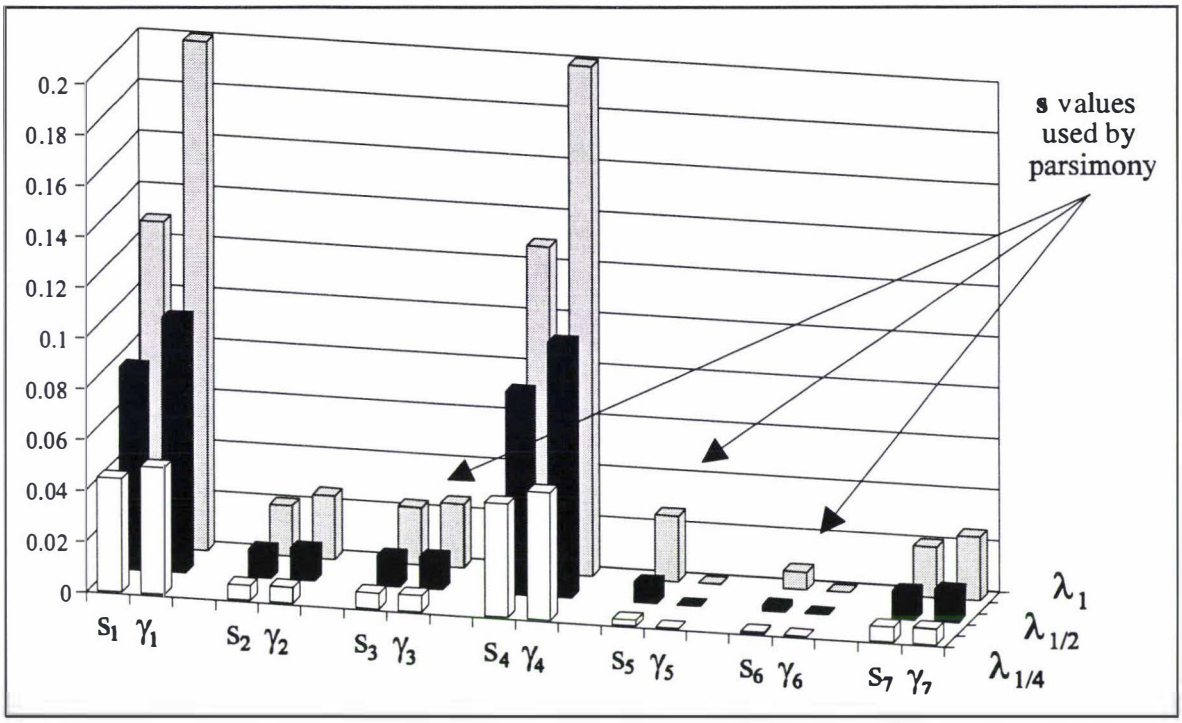


FIGURE 4.1 Vectors $\mathbf{s}(T)$ and $\boldsymbol{\gamma}(T)$ derived from the tree in figure 2.1 with three scalings of edge weights; 1/4 (white = $\lambda_{1/4}$ in text), 1/2 (dark grey = $\lambda_{1/2}$) and 1 (light grey = λ_1) ($s(T)_0$ and $\gamma(T)_0$ omitted). The y axis measures s_i as the observed proportion of sites showing that pattern, while γ_i is measured in expected number of substitutions per site on the edges of the tree. Note the relative size of s_i to γ_i for the parsimony "informative" i.e. non-pendant bipartitions, s_3, s_5 and s_6. Bipartition s_5 exceeds the size of s_3 when the rate of change per site (λ) is highest, indicating that except for sampling error, the parsimony tree selection criteria applied to $\mathbf{s}(T)$ will choose the wrong tree. Parsimony will however be consistent when applied to $\boldsymbol{\gamma}$ (e.g. Steel *et al.* 1993b).

4.2.2 The variance-covariance matrix $\mathbf{V}[\hat{\mathbf{s}}]$ of the sequence spectrum $\hat{\mathbf{s}}$

The entries in the multinomial population variance-covariance matrix $\mathbf{V}[\hat{\mathbf{s}}] = V_{ij}$ (referred to from now as a covariance matrix) of $\hat{\mathbf{s}}$ estimated from c sequence sites are

$$V_{ii} = s(T)_i(1-s(T)_i) / c \,, \qquad (4.2.2\text{-}1)$$

$$V_{ij} = -s(T)_i s(T)_j / c, \qquad (4.2.2\text{-}2)$$

where V_{ii} is the variance of $\hat{s}_i$ and V_{ij} ($i \neq j$) is the covariance of $\hat{s}_i$ with $\hat{s}_j$. $\mathbf{V}[\hat{\mathbf{s}}]$ is symmetric with $(2^{t-1})^2 = 4^{t-1}$ entries. Table 4.1 shows the covariance matrix of our example. Note: For ease of presentation, none of the values that appear in covariance matrices of this section are divided by the sequence length. With sampled data we replace $\mathbf{s}(T)$ with its maximum likelihood estimate $\hat{\mathbf{s}}$, obtaining the sample covariance matrix $\hat{\mathbf{V}}[\hat{\mathbf{s}}]$.

The marginal distributions of entries in $\hat{\mathbf{s}}$ are binomial distributions (with parameters c and s_i) scaled by 1/c. When cs_i and $c(1-s_i)$ are both > 5 then the normal approximation to the binomial is reasonable (Freund and Walpole 1987 p. 229). A better approximation to these binomials when c is large and cs_i is less than 5 is the Poisson distribution. When all $cs_i > 5$ then $\hat{\mathbf{s}}$ is distributed approximately multivariate normally. As $c \to \infty$, $\hat{\mathbf{s}}$ tends to a multivariate normal distribution with variance tending to zero.

Table 4.1 The vector s(T) and the covariance matrix V[$\hat{\mathbf{s}}$]

Index	0	1	2	3	4	5	6	7
Bipartition	{0/123}	{1/234}	{2/134}	{12/34}	{3/124}	{13/24}	{23/14}	{123/4}
s(T) →	.6479	.1283	.0200	.0226	.1283	.0258	.0070	.0200
V[$\hat{\mathbf{s}}$] Index								
0	.2281	-.0831	-.0130	-.0146	-.0831	-.0167	-.0046	-.0130
1	-.0831	.1119	-.0026	-.0029	-.0165	-.0033	-.0009	-.0026
2	-.0130	-.0026	.0196	-.0005	-.0026	-.0005	-.0001	-.0004
3	-.0146	-.0029	-.0005	.0221	-.0029	-.0006	-.0002	-.0005
4	-.0831	-.0165	-.0026	-.0029	.1119	-.0033	-.0009	-.0026
5	-.0167	-.0033	-.0005	-.0006	-.0033	.0251	-.0002	-.0005
6	-.0046	-.0009	-.0001	-.0002	-.0009	-.0002	.0070	-.0001
7	-.0130	-.0026	-.0004	-.0005	-.0026	-.0005	-.0001	.0196

Note: Vector **s**(T) was generated form the vector γ(T) in figure 2.1(B) using equation 4.2.1-1. The indexing of the bipartitions is that of Hendy and Penny (1993). Following this is the multinomial covariance matrix of $\hat{\mathbf{s}}$, **V**[$\hat{\mathbf{s}}$]. For convenience, the values in this and the subsequent covariance matrices (tables 4.2, 4.3 and 4.4) are calculated independent of a sequence length, so, to calculate their entries for a sequence of length *c*, simply divide each entry by c.

For illustrative purposes the inverse of the Hadamard conjugation

$$\hat{\boldsymbol{\gamma}} = \mathbf{H}^{-1}(\ln(\mathbf{H}\hat{\mathbf{s}})), \tag{4.2.2-3}$$

is broken up into three parts,

$$\hat{\mathbf{r}} = \mathbf{H}\hat{\mathbf{s}}; \tag{4.2.2-4}$$

$$\hat{\boldsymbol{\rho}} = \ln \hat{\mathbf{r}}; \tag{4.2.2-5}$$

$$\hat{\boldsymbol{\gamma}} = \mathbf{H}^{-1}\hat{\boldsymbol{\rho}}, \tag{4.2.2-6}$$

which will be analysed in turn.

4.2.3 The calculation of $\hat{\mathbf{r}}$ (= $\mathbf{H}\hat{\mathbf{s}}$) and its covariance matrix V[$\hat{\mathbf{r}}$]

The first operation in the Hadamard conjugation is the calculation of $\hat{\mathbf{r}}$. The Hadamard matrix, **H**, is a matrix of plus and minus ones. A subset of the entries in $\hat{\mathbf{r}}$ are $1 - 2d_{ij}$, where d_{ij} is the observed distance between taxon i and taxon j. Since $\mathbf{r} = \mathbf{H}\mathbf{s}$ is a linear transformation

$$\mathbf{V}[\hat{\mathbf{r}}] = \mathbf{H}\mathbf{V}\mathbf{H}^{t}, \tag{4.2.3-1}$$

where **V** is the covariance matrix of $\hat{\mathbf{s}}$ (Krzanowski 1988, p. 205). (See Appendix 4.1 for a more efficient implementation of these operations). Table 4.2 shows the results of these operations on **V**[$\hat{\mathbf{s}}$].

The entries in $\hat{\mathbf{r}}$ (like the entries in $\hat{\mathbf{s}}$) have marginal distributions which are scaled binomials even though their covariance structure is not that of a multinomial. We now describe the marginal distributions of entries in $\hat{\mathbf{r}}$:

For any subset $J \subset \{0, ..., 2^{t-1} - 1\}$ of indices, let $\hat{f}_J = \sum_{j \in J} \hat{f}_j$ and $s_J = \sum_{j \in J} s_j$ (where s_j is from s(T)).

Recalling that $\hat{\mathbf{f}}$ ($= c\hat{\mathbf{s}}$), then $\hat{f}_J$ is distributed as a binomial with parameters c and s_J i.e. ~ B(c, s_J)(that is binomial with parameters c and s_J). Since $\hat{\mathbf{s}} = \hat{\mathbf{f}} / c$, then $\hat{s}_J$ will be distributed as a scaled binomial, that is $\hat{s}_J$ ~ (B(c, s_J) scaled by 1/c). Let $J_i = \{j: h_{ij} = -1\}$(h from the Hadamard matrix), then $\hat{r}_i = 1 - 2\hat{s}_{J_i}$ is distributed as 1 - 2 (B(c, s_{J_i}) scaled by 1/c). The variance of $\hat{r}_i$ is therefore $4s_{J_i}(1 - s_{J_i})/c = (1-r_i^2)/c$. In biological data sets entries in $\hat{\mathbf{r}}$ will often be estimated from an expected 5 or more observable sequence changes i.e. $E[\hat{f}_j]$ is 5 or more. Correspondingly the marginal distributions of $\hat{r}_i$ will often be well approximated a normal distribution with mean r_i and variance as given above.

Since $\hat{\mathbf{r}}$ is a linear transformation of $\hat{\mathbf{s}}$, then both $\hat{\mathbf{s}}$ and $\hat{\mathbf{r}}$ tend to multivariate normal as sequences get longer ($c \rightarrow \infty$), with variance tending to zero. This is a generalisation of the pairwise distance matrix (a subset of $\hat{\mathbf{r}}$) tending to multivariate normal (e.g. Bulmer 1991a).

Table 4.2 The covariance matrix of $\hat{\mathbf{r}}$, $V[\hat{\mathbf{r}}]$.

Index	0	1	2	3	4	5	6	7
Pathset	{0}	{1,4 }	{2,4}	{1,2}	{3,4}	{1,3}	{2,3}	{1,2; 3,4}
r(T)	1.000	.6065	.8607	.6376	.6376	.4274	.6065	.4066
$\mathbf{V}[\hat{\mathbf{r}}]$								
Index								
0	.0000	.0000	.0000	.0000	.0000	.0000	.0000	.0000
1	.0000	.6321	.1156	.4740	.0407	.3784	.0387	.3599
2	.0000	.1156	.2592	.0577	.0577	.0387	.1156	.0775
3	.0000	.4740	.0577	.5934	.0000	.3340	.0407	.3784
4	.0000	.0407	.0577	.0000	.5934	.3340	.4740	.3784
5	.0000	.3784	.0387	.3340	.3340	.8173	.3784	.6869
6	.0000	.0387	.1156	.0407	.4740	.3784	.6321	.3599
7	.0000	.3599	.0775	.3784	.3784	.6869	.3599	.8347

Note: The first row and column of this matrix will always be 0 since r_0 the sum of all entries in $\hat{\mathbf{s}}$ is 1. For convenience of presentation, all entries in $V[\hat{\mathbf{r}}]$ have yet to be divided by the sequence length, c.

4.2.4 The covariance matrix $V[\hat{\rho}]$ of the estimated path lengths $\hat{\rho}$

The total implied number of changes per site on the *i*-th set of edge paths is $-\hat{\rho}_i / 2$, where $\hat{\rho}_i = \ln \hat{r}_i$. This value is also the componentwise maximum likelihood estimate of ρ_i (this follows, since $\hat{r}_i$ is the ML estimator of $r(T)_i$, and the value of $\rho(T)_i$ most likely to have given $\hat{r}_i$ is just $\rho(T)_i = \ln(r(T)_i)$. The estimates of corrected pathset length are expressed as minus twice their expected value, since multiplication of $\hat{\rho}$ by $\mathbf{H}^{-1}$ will then give the implied total number of changes on each bipartition in $\boldsymbol{\gamma}$, the corrected sequence vector.

The entries in $\mathbf{V}[\hat{\mathbf{r}}]$ must be adjusted for this non-linear transformation to give $\mathbf{V}[\hat{\rho}]$. We may estimate the variance of $\hat{\rho}_i$ exactly as var $[\hat{\rho}_i]$ = var $[\ln(\hat{r}_i/c)]$ which involves summing

many binomial terms. This computation is time consuming so we estimate the variance of ρ_i using the first order approximation known as the delta method (Stuart and Ord 1987, p. 324). The delta method uses the result that any transformation changes the variance by a factor approximately equal to the square of the gradient of the transformation curve. We denote covariance and correlation matrices estimated using the delta method with a prime e.g. $V'[\hat{\rho}]$. The gradient of the transformation $\rho_i = \ln(r_i)$ with respect to r_i is $1/r_i$. So letting Vij be the ij-th entry in $\mathbf{V}[\hat{\mathbf{r}}]$ we have

$$V'[\hat{\rho}]_{ii} = (1/r_i)^2 V_{ii} \qquad (4.2.4\text{-}1)$$

a similar result for covariances gives

$$V'[\hat{\rho}]_{ij} = 1/r_i.1/r_j V_{ij}, \qquad (4.2.4\text{-}2)$$

where r_i and r_j are respectively the *i*-th and *j*-th entries in the vector $\mathbf{r}$ and $i \neq j$. Table 4.3 shows the results of these calculations. Later we investigate the size and direction of bias when using the delta method.

All values in $\hat{\rho}$ will have greater variances and covariances than the corresponding entries in $\hat{\mathbf{r}}$ because the gradient of the log curve with respect to the r axis is 1 at $r_i = 1$, increasing without limit as r_i tends to 0. As $c \rightarrow \infty$, $\hat{\mathbf{r}}$ tends to multivariate normality with variance $\rightarrow 0$, so $\hat{\rho}$ also tends to multivariate normality with variance $\rightarrow 0$. The logarithmic function introduces positive skewness to the marginal distributions of $\hat{\rho}$ relative to $\hat{\mathbf{r}}$. This skewness (which causes bias since $E[\ln(\hat{r}_i)]$ is always > $\ln(E[\hat{r}_i])$, E being the expected value) tends to zero as *c* tends to infinity (and variance goes to zero). Skewness also tends to zero as λ (the rate of change per site) goes to zero (i.e. the paths require almost no correction).

If pathsets *k* and *l* have no edges in the tree in common then their correlation will be zero. Generally, for low rates of change, the covariance of two additive pathsets will be approximately the sum of the variances of the lengths of the edges they share in common. This is a standard result of linear variables; that is if $g = \sum_{i=1}^{k} a_i x_i$, where a_i is a scalar, and x_i is a random variable, then $\text{var}[g] = \sum a_i^2 \text{var}[x_i] + \sum_{i \neq j} a_i a_j \text{cov}[x_i, x_j]$ (Stuart and Ord 1987, p. 324). This result is obvious when it is noted that in this instance the a_i's are all 1, and the covariances between edges at low rates tend to zero. Further because there is practically no correction for multiple substitutions at low rates, the mean and the variance of an edge *i* converges to the mean and the variance of the pattern s_i. In turn, pattern s_i has a scaled binomial marginal distribution, which converges to a Poisson distribution with variance = s_i / c. Therefore, the correlation of pathset *k* and *l* (at low rates of change) is approximately the sum of edges in common, divided by the square root of the length of pathset *k* then divided by the square root of the length of pathset *l* (since cor[k,l] = cov[k,l] / (s.d.[k] × s.d.[l]) , and the effect of *c* on the top and the bottom line of this equation cancels). So, as $\hat{\rho}_k$ and $\hat{\rho}_l$ count events on increasingly equivalent pathsets, their covariance tends to their variance, and their correlation tends to one. This interpretation is a

generalisation of the covariance structure of pairwise distances on trees (Nei and Jin 1989, Bulmer 1991a). Unfortunately this relationship breaks down at higher rates of change, such as the example used here.

Table 4.3 The covariance matrix V'[$\hat{\rho}$] estimated using the delta method

Index	0	1	2	3	4	5	6	7
Pathset	{0}	{1,4 }	{2,4}	{1,2}	{3,4}	{1,3}	{2,3}	{1,2; 3,4}
ρ(T)	.00	-.50	-.15	-.45	-.45	-.85	-.50	-.90
V'[$\hat{\rho}$]								
Index								
0	.000	.000	.000	.000	.000	.000	.000	.000
1	.000	1.718	.221	1.226	.105	1.460	.105	1.460
2	.000	.221	.350	.105	.105	.105	.221	.221
3	.000	1.226	.105	1.460	.000	1.226	.105	1.460
4	.000	.105	.105	.000	1.460	1.226	1.226	1.460
5	.000	1.460	.105	1.226	1.226	4.474	1.460	3.953
6	.000	.105	.221	.105	1.226	1.460	1.718	1.460
7	.000	1.460	.221	1.460	1.460	3.953	1.460	5.050

To obtain pathset lengths from ρ(T), multiply entries by -2; to obtain the variances of these pathset lengths divide entries in the covariance matrix by 4. The variances of the longest sets of paths, $\hat{\rho}_5$ (-2 the estimated distance between taxa 1 and 3) and $\hat{\rho}_7$ (-2 the estimated length of the non-intersecting pathset between taxa 1, 2, 3 and 4 (i.e edges 1, 2, 4 and 7)), have increased considerably due to the uncertainty created by multiple substitutions at a site. Since these two entries estimate the weights on an very similar sets of tree edges (which equates to counting mostly identical sequence bipartitions) they have a large covariance and a correlation of approximately 0.85 (cor(i,j) = cov(i,j) / [s.d.(i) × s.d.(j)]).

4.2.5 The covariance and correlation matrix of $\hat{\gamma}$, the corrected data

The conjugate spectrum $\hat{\gamma}$ of transformed bipartitions is calculated from $\hat{\rho}$ by equation 4.2.2-6. Its covariance matrix $\mathbf{V}[\hat{\gamma}]$ is equal (Krzanowski 1988, p. 205) to

$$\mathbf{V}[\hat{\gamma}] = \mathbf{H}^{-1}\mathbf{V}[\hat{\rho}]\mathbf{H}^{-1}. \qquad (4.2.5\text{-}1)$$

These operations are computationally almost identical to those involved in going from $\mathbf{V}[\hat{\mathbf{s}}]$ to $\mathbf{V}[\hat{\mathbf{r}}]$ since $\mathbf{H}^{-1}$ is $(1/2^{t-1})\mathbf{H}$. Alternatively we can describe the whole series of transformations as $\mathbf{V}[\hat{\gamma}] = \mathbf{H}^{-1}(\delta(\mathbf{H}\mathbf{V}[\hat{\mathbf{s}}]\mathbf{H}))\mathbf{H}^{-1}$, where δ is the delta method applied as described in equations (4.2.4-1) and (4.2.4-2). We still have the same 2^{t-1}-1 degrees of freedom in $\hat{\gamma}$ that we started with in $\hat{\mathbf{s}}$.

Table 4.4 shows $\mathbf{V}'[\hat{\gamma}]$, while figure 4.2 shows the relative sizes of errors on entries in $\hat{\gamma}$ as implied by $\mathbf{V}'[\hat{\gamma}]$ with c = 1000. Hendy and Penny (1989) observed that long edges cause parsimony on uncorrected data to be inconsistent (paraphrased as "long edges attract"). Here we observe that with a consistent method of tree building, long edges cause increased variance of corrected bipartitions which are not in the tree that generated the data. In this case (and we expect generally with appropriate path corrections) there is little increase in the variance of the

entry in $\hat{\gamma}_i$ relating to the correct internal edge(s) but a large increase in the variance of entries in $\hat{\gamma}$ that group these long edges together. This will greatly increase the probability (with respect to the same tree without long pendant edges) of such non-tree entries being in the tree selected by any optimality criteria, that is choose the wrong tree. Generally, resolving the branching order around long edges will be more error prone and require more data for the same degree of precision than if all edges were the same length. This result is implied for other methods as well e.g. ML (more on this in chapter 5).

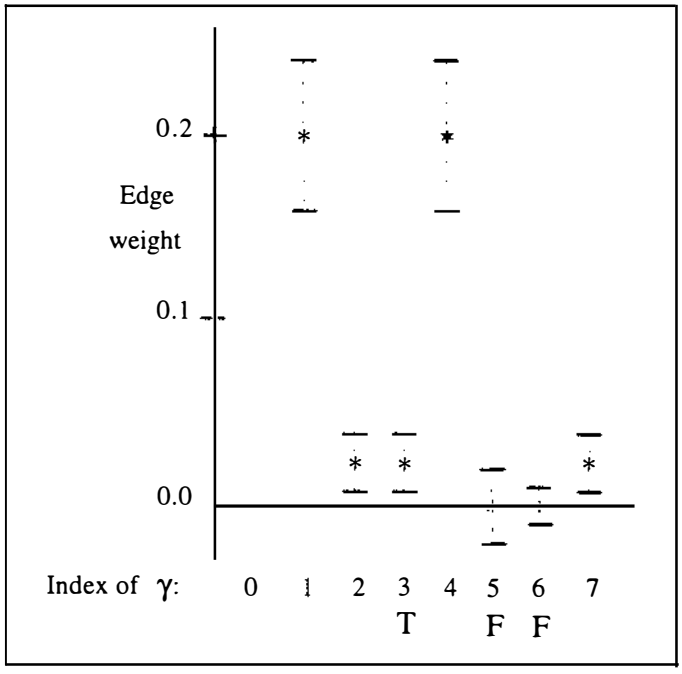


FIGURE 4.2 Error bars for the estimator $\hat{\gamma}$ made with random samples of $c = 1000$ sites from $\mathbf{s}(T)$ (fig 4.2, grey columns)(γ_0 excluded). Even for a sequence of this length there is significant overlap in the distribution of $\hat{\gamma}_3$ (representing support for the internal edge in the tree generating the data, labeled with a T) and $\hat{\gamma}_5$ (support for an erroneous tree, labeled with an F). The magnitude of the size of error on $\hat{\gamma}_i$ correlates highly with the variance of $\hat{s}_i$ (see text).

Table 4.4 The covariance matrix V'[$\hat{\gamma}$] (via the delta method)

Index	0	1	2	3	4	5	6	7
Bipartition	{0/1234}	{1/234}	{2/134}	{12/34}	{3/124}	{13/24}	{23/14}	{123/4}
$\gamma(T)\rightarrow$	.475	.200	.025	.025	.200	.000	.000	.025
V'[$\hat{\gamma}$]								
Index								
0	.8445	-.4318	-.0389	-.0250	-.4318	.1022	.0197	-.0389
1	-.4318	.4336	.0020	-.0039	.1016	<u>-.1019</u>	-.0112	.0116
2	-.0389	.0020	.0604	-.0235	.0116	-.0039	-.0161	.0084
3	-.0250	-.0039	-.0235	.0684	-.0039	.0036	.0077	-.0235
4	-.4318	.1016	.0116	-.0039	.4336	<u>-.1019</u>	-.0112	.0020
5	.1022	<u>-.1019</u>	-.0039	.0036	<u>-.1019</u>	.1031	.0027	-.0039
6	.0197	-.0112	-.0161	.0077	-.0112	.0027	.0246	-.0161
7	-.0389	.0116	.0084	-.0235	.0020	-.0039	-.0161	.0604

The largest covariance (excluding $\hat{\gamma}_0$ which cannot be an edge in any tree) is the negative covariance (underlined) between either $\hat{\gamma}_1$ and $\hat{\gamma}_5$, or $\hat{\gamma}_4$ and $\hat{\gamma}_5$. That is, between either long pendant edge 1 or 4 of the model tree (fig. 2.1), and $\hat{\gamma}_5$ which groups these long edges together. Bipartition s_5 causes parsimony applied to $\mathbf{s}(T)$ to converge to the wrong tree, while $\hat{\gamma}_5$ (its corrected estimated value) has a relatively large variance. This larger variance makes $\hat{\gamma}_5$ more likely than $\hat{\gamma}_6$ to be greater than $\hat{\gamma}_3$ (which estimates the internal edge in our model tree) by chance. Many optimality criteria such as parsimony, compatibility

or closest tree when applied to this $\hat{\gamma}$ will simply choose the largest of the three previous entries as the internal edge in the optimal tree. For ease of presentation values in this matrix have yet to be divided by c.

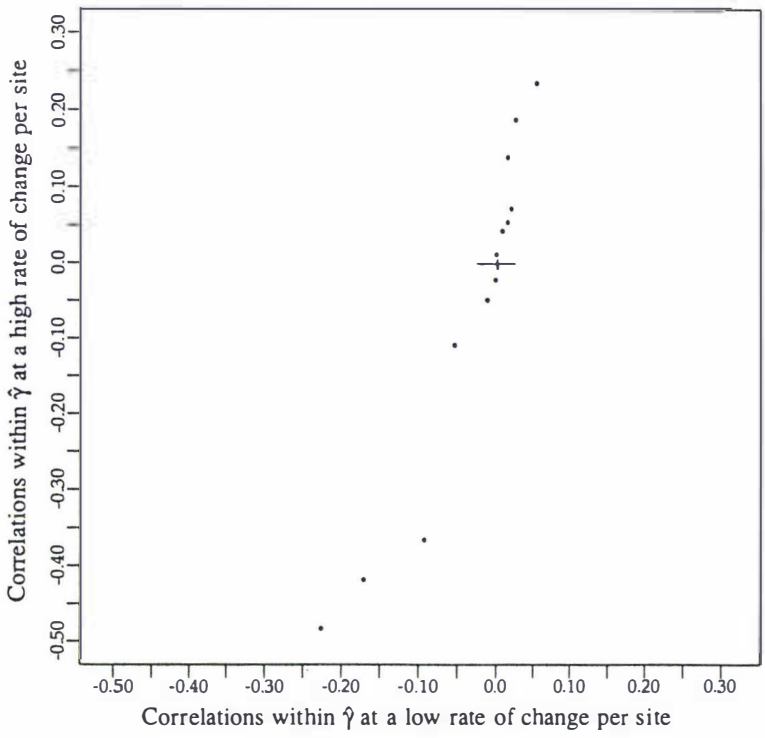


FIGURE 4.3 A plot of the correlations of pairs of entries in $\hat{\gamma}$ ($\hat{\gamma}_0$ excluded) at a low rate of change per site ($\lambda_{1/4}$) on the x-axis vs the corresponding correlations within $\hat{\gamma}$ at a high rate of change (λ_1) (see text). The bar marks the origin and the axes are to scale. The trend in this graph illustrates how higher rates of change per site lead to higher correlations within $\hat{\gamma}$. This effect is due to higher rates of change causing more multiple changes. Calculated with c = 1000.

Table 4.5 The correlation matrix C'[$\hat{\gamma}$] (via the delta method)

Index	0	1	2	3	4	5	6	7
Bipartition	{0/1234}	{1/234}	{2/134}	{12/34}	{3/124}	{13/24}	{23/14}	{123/4}
$\gamma(T) \rightarrow$	-.475	.200	.025	.025	.200	.000	.000	.025
C'[$\hat{\gamma}$] Index								
0	1.000	-.713	-.172	-.103	-.713	.346	.136	-.172
1	-.713	1.000	.012	-.022	.234	<u>-.482</u>	-.108	.071
2	-.172	.012	1.000	-.365	.071	-.048	-.418	.138
3	-.103	-.022	-.365	1.000	-.022	.042	.187	-.365
4	-.713	.234	.071	-.022	1.000	<u>-.482</u>	-.108	.012
5	.346	<u>-.482</u>	-.048	.042	<u>-.482</u>	1.000	.053	-.048
6	.136	-.108	-.418	.187	-.108	.053	1.000	-.418
7	-.172	.071	.138	-.365	.012	-.048	-.418	1.000

Note: The largest correlations between entries in $\hat{\gamma}$ (excluding $\hat{\gamma}0$) are the negative correlations (underlined) between γ_1 or γ_4 (estimates of the long pendant edges 1 and 4 respectively) and γ_5 which corresponds to the "parsimony goes wrong pattern" s_5 that groups the long edges together.

Let $\mathbf{C}[\hat{\gamma}]$ be the correlation matrix of $\hat{\gamma}$ so

$$\mathbf{C}'[\hat{\gamma}] = \mathbf{W}\mathbf{V}'[\hat{\gamma}]\mathbf{W}, \qquad (4.2.5\text{-}2)$$

where $\mathbf{W}$ is a diagonal matrix with entries $W_{ii} = 1/\sqrt{V_{ii}}$ (the inverse of the standard deviations of $\hat{\gamma}$). The correlations of entries in $\hat{\gamma}$ (excluding those with $\hat{\gamma}_0$) for the tree in figure 2.1a and with $c = 1000$ are apparent as the ellipsoid bivariate distributions shown in figure 4.5b.

We consider now the causes of the correlation structure of the weighted transformed bipartitions ($\hat{\gamma}_i$, $i \neq 0$). All entries in $\hat{\gamma}$ are correlated. Excluding $\hat{\gamma}_0$, these range in value from -0.48 to +0.23. This is due to the non-independence of cells in the multinomially distributed $\hat{\mathbf{s}}$ and the transformation for multiple changes which alter the relative sizes of these correlations. For example the cells in $\hat{\mathbf{s}}$ excluding the larger cells s_0, s_1, and s_4 have correlations between 0 and -0.07 (not shown), but the corresponding small entries in $\hat{\gamma}$ may have large correlations, for example, $\hat{\gamma}_2$ and $\hat{\gamma}_6$ have a correlation of -0.46.

The differences between corresponding entries in $\mathbf{C}[\hat{\gamma}]$ and $\mathbf{C}[\hat{\mathbf{s}}]$ (the multinomial correlation matrix of $\hat{\mathbf{s}}$) are the result of multiple changes at sites in sequences evolving on the weighted tree in fig. 1. For example in table 4.5 the largest correlation is the negative correlation (-0.48) of $\hat{\gamma}_5$ with both $\hat{\gamma}_1$ and $\hat{\gamma}_4$. This correlation is predominantly due to parallel changes along the edges $\gamma(T)_1$ and $\gamma(T)_4$. The majority of such parallel changes result in patterns s_1 or s_4 becoming pattern s_5, which adds support to $\hat{\gamma}_5$ at the expense of $\hat{\gamma}_1$ and $\hat{\gamma}_4$. The only other double change that would add support to s_5 would be parallel changes on the short edges e_2 and e_7. This is much less likely, as are sets of 3 or more changes giving the pattern s_5. As a result the small negative correlations between $\hat{\gamma}_5$ and either $\hat{\gamma}_2$ or $\hat{\gamma}_7$ are little changed from the corresponding entries in $\mathbf{C}[\hat{\mathbf{s}}]$. The impact of multiple substitutions in increasingly differentiating the correlation structure of $\hat{\gamma}$ from the multinomial correlations of $\hat{\mathbf{s}}$ is illustrated in figure 4.3. In this plot the only difference between axes is that λ (the rate of change per site) on the x axis is only 1/4 of that on the y axis (that is the correlations of the white vs the light grey columns in fig. 2). As $\lambda \rightarrow 0$ the correlations of $\hat{\gamma}$ ($\hat{\gamma}_i$, $i \neq 0$) tend to those of $\hat{\mathbf{s}}$, which also tend to zero.

4.3 THE MARGINAL DISTRIBUTIONS OF ENTRIES IN $\hat{\gamma}$

In order to know the probability of entry $\hat{\gamma}_i$ exceeding a certain value due to sampling error alone we need to know not only its mean and variance but also its marginal distribution. Here we explore the marginal distributions of entries in $\hat{\gamma}$ by varying:

(1) Sequence lengths (using c = 100 and 1000).

(2) The rate of change per site (λ) without altering the relative weights of edges on the tree.

We denote the relatively high rate in our model tree (fig. 1a) as λ_1, then reduce this rate by a factor of two ($\lambda_{1/2}$) and four ($\lambda_{1/4}$) as shown in figure 4.1.

We will restrict our attention to the entries in $\hat{\gamma}$ which may be used as edge length estimates in a tree and ignore γ_0, which is a function of the other 7 entries. Due to the symmetry of this tree, the Monte Carlo distributions of $\hat{\gamma}_1$ are equivalent to those of $\hat{\gamma}_4$ and likewise for the pair $\hat{\gamma}_2$ and $\hat{\gamma}_7$. All combinations of c and λ were checked with high resolution univariate, bivariate and normal probability plots but only illustrative examples are shown here.

The generalities that emerge are:

(1) An arbitrarily small rate of change implies few multiple changes so that the conjugation causes little difference between the values in $\hat{\mathbf{s}}$ and $\hat{\gamma}$. Consequentially $\hat{\gamma}$ has a distribution close to a multinomial. Figures 4.4a and 4.4b illustrate the binomial-like marginal distributions of $\hat{\gamma}_1$ and $\hat{\gamma}_3$ with $\lambda_{1/4}$ and c = 100. Increasing c increases the expected values in $\hat{\mathbf{s}}$ causing the marginal distributions to become more normal in shape.

(2) When c is small and $\hat{\mathbf{s}}$ is distributed like a multinomial there may be slightly non-linear relationships between entries in $\hat{\gamma}$ irrespective of the rate of change (for example $\hat{\gamma}_1$ and $\hat{\gamma}_5$ in figure 4.5a, c = 100 and $\lambda_{1/4}$). High rates of change can exacerbate such effects. These relationships between $\hat{\gamma}_i$ and $\hat{\gamma}_j$ become linear as c increases. Entries in $\hat{\gamma}$ with c = 1000 and λ_1 (fig. 4.5b) show the distribution has converged close to multivariate normality (apart from a slight amount of skewness).

(3) Smoothing due to correction for implied multiple changes can introduce either positive or negative skewness to the marginal distributions entries in $\hat{\gamma}$. For example the bipartition $\hat{\gamma}_5$ in fig. 5d shows smoothing and skewness. Smoothing has erased the discreteness present in the marginal distribution of $\hat{s}_5$ (~ B(100, 3.6)). Smoothing is pronounced on this variable because the two longest sets of paths $\hat{\rho}_5$ and $\hat{\rho}_7$, which have the highest skewness, variance, and also a high correlation, both have the same sign in the summation from ρ to give γ_5. That is, the largest potential errors after taking into account multiple substitutions are additive when estimating $\hat{\gamma}_5$. The entries in $\hat{\gamma}$ relating to the two longest tree edges (γ_1 and γ_4) have skewness and smoothing for the same reason, but their skewness is positive due to a reversal of the signs on $\hat{\rho}_5$ and $\hat{\rho}_7$ in their estimation from ρ (fig. 4.4c).

(4) High λ and small c both increase the skewness resulting from correction for multiple changes, which in turn results in bias. Figures 4.4c and d show the skewness on $\hat{\gamma}_1$ and $\hat{\gamma}_5$ with $c = 100$ and λ_1. Increasing c to 1000 as shown in fig. 4.5b reduces the skewness greatly.

(5) If other factors are kept constant, as c increases the amount of smoothing decreases around individual peaks in the marginal distributions of $\hat{\gamma}$. However, this decrease is not as rapid as the decrease in the size of the gaps between peaks, so the net result is that the marginal distributions of $\hat{\gamma}$ become effectively continuous as c becomes large.

(6) In general if the variance of $\hat{\gamma}_i$ is greater than 5/c then its marginal distribution approximates a normal distribution in overall shape. Figure 4.4a illustrates a worst case example where $\mathrm{Var}[\hat{\gamma}_i]$ just meets this criterion, and the marginal distribution retains much binomial character. When the variance of $\hat{\gamma}_i$ is less than 5/c it tends to show the characteristics of a binomial distribution with an expected value of less than 5. An example of this is seen in figure 4.4b. We make these comments in light of the commonly used rule for approximating a binomial with a normal distribution when cs_i and $c(1-s_i)$ are > 5 (Freund and Walpole 1987 p. 229). Chapter 6 describes hypothesis testing of entries in $\hat{\gamma}$, elaborating upon these findings.

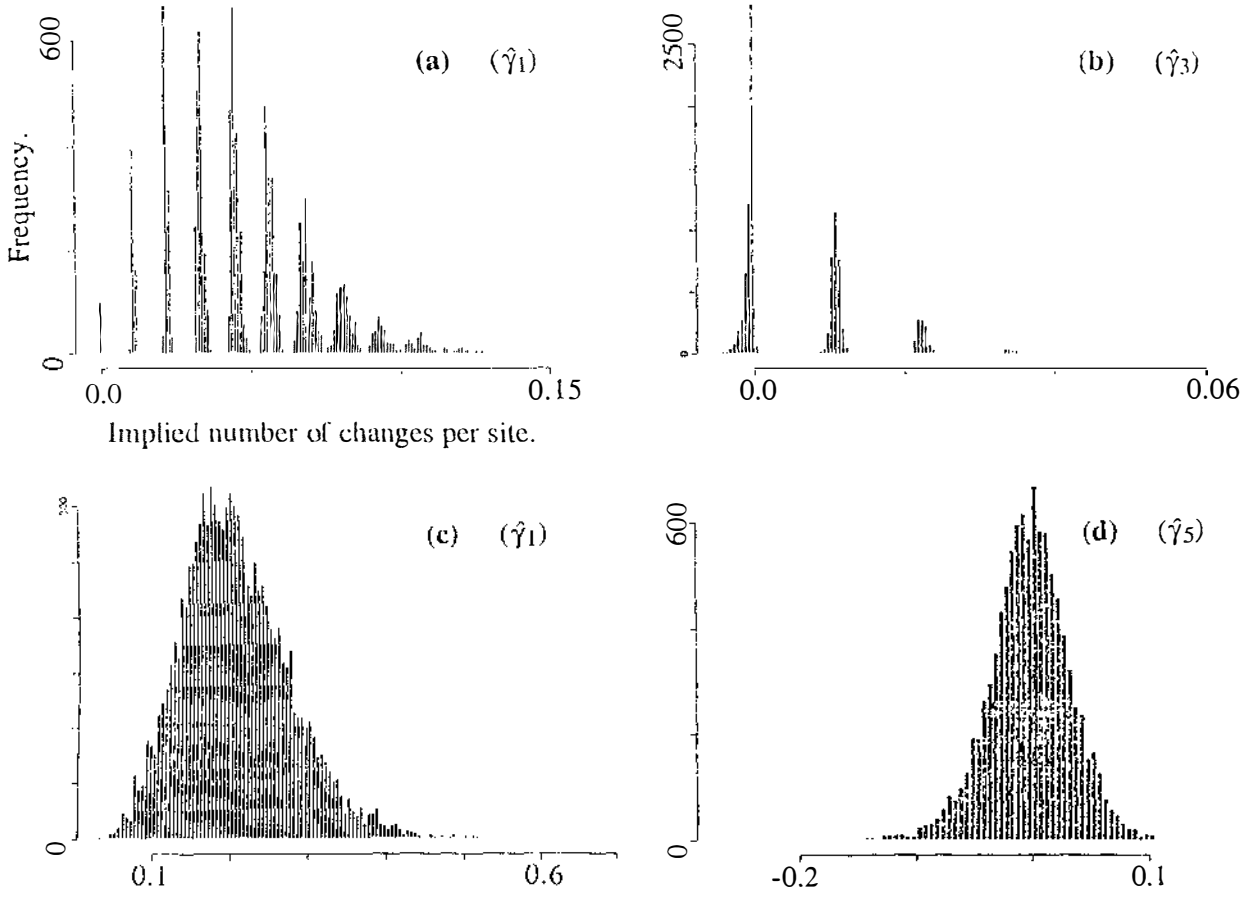


FIGURE 4.4 Marginal distributions of $\hat{\gamma}$ from a simulation of size n = 10, 000. **4.4a** and **4.4b** $\hat{\gamma}_1$ and $\hat{\gamma}_3$ with $\lambda_{1/4}$ and c = 100. **4.4c** and **4.4d** $\hat{\gamma}_1$ and $\hat{\gamma}_5$ with c = 100 and λ_1.

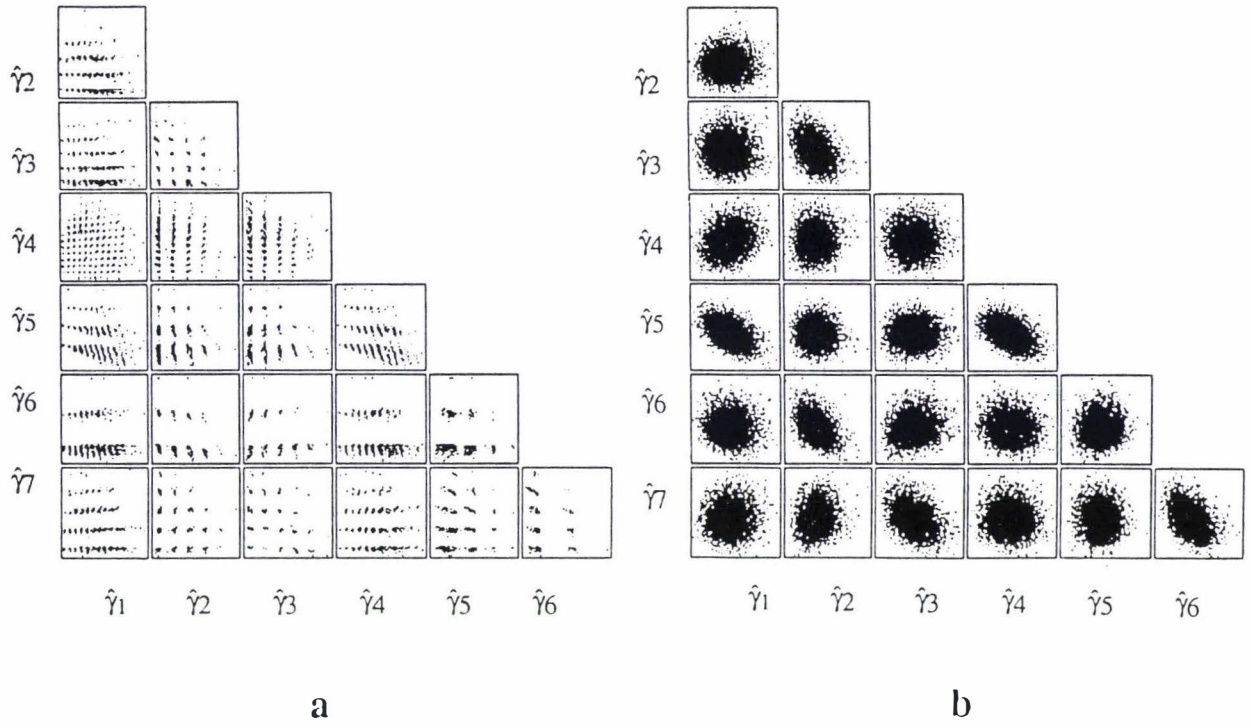


FIGURE 4.5 Bivariate plots of entries in $\hat{\gamma}$ (excluding $\hat{\gamma}_0$) from a simulation of size n = 10,000. **4.5a** Parameters c = 100 and $\lambda_1/4$. **4.5b** Parameters c = 1000 and λ_1.

In these simulations the largest bias occurred on variable $\hat{\gamma}_1$ with c = 100 and λ_1. This bias was 0.0065 compared to an expected mean of 0.2, which contributed less than 10% to the mean square error (MSE) of estimating $\gamma(T)_1$ from $\hat{\gamma}_1$ (MSE = bias^2 + variance, which here are 4×10^{-5} and 43×10^{-5} respectively). Accordingly we expect that the bias in the Hadamard conjugation due to the log transform is likely to have minor bearing upon estimated edge lengths in trees calculated from sampled data. This should be especially true with real data where the choice of model, and therefore corrections for multiple changes, are likely to be a much larger source of error. This bias could be eliminated with an unbiased estimator of inferred path lengths to replace the maximum likelihood estimator used here (equivalent to replacing n with n-1 when estimating sample variances).

Simulations (including those on other trees, not shown) and theory suggest that the two factors c and λ will influence the form of the marginal distributions of bipartitions generated via the Hadamard conjugation in the ways illustrated above, irrespective of the number of taxa or tree shape. The major effect of tree shape (edge lengths plus topology) is that parallel changes are most likely on long edges. When long edges are separated by short internal edges, the variances of the $\hat{\gamma}_i$ which suggest grouping the long edges together are often of a similar size to the variances of the $\hat{\gamma}_i$ for the correct internal edges. For short sequences (where the size of the standard error of the $\hat{\gamma}_i$ relating to a short internal edge is of similar magnitude to its expected value), these factors can make it quite likely that any tree selection criteria applied to $\hat{\gamma}$ will choose the wrong $\hat{\gamma}_i$ when resolving such a short internal edge(s). We also expect that, in general, the marginal distributions of $\hat{\gamma}_i$ relating to $\hat{s}_i$ which are the result of relatively common parallel changes will show the same sort of smoothing that we saw occurring to the marginal

distribution of $\hat{\gamma}_5$ in the above simulations. The form of the marginal distributions in these cases will, however, remain of secondary importance to the absolute size of standard errors in determining the size range that a $\hat{\gamma}_i$ is expected to take due to sampling error.

The other major effect upon the distribution of $\hat{\gamma}$ is that, as the number of taxa (t) increases, the number of possible sequence patterns (bipartitions with 2-states, generically labeled quadrapartitions with 4-states) increases exponentially. In addition most of these patterns will be incompatible with any given tree, which typically results in any such pattern having a low expected frequency of occurrence. So for even modest t the majority of entries in $\hat{\mathbf{s}}$ will have expected frequencies of < 1 for any biologically realistic sequence length (even for sequences of the order of 10^6). Consequentially, most entries in $\hat{\gamma}$ will have marginal distributions like those of a Poisson variable with an expected values of < 1. This poses a major problem for statistical testing, especially for overall tests of fit of data to model (including the tree), where the most tractable approaches using multivariate normal statistics (e.g. Bulmer 1991a) will often be invalid. In chapter 6 we consider strategies to negate this problem.

4.4 PROPERTIES OF DELTA METHOD COVARIANCE MATRICES

In this section, the properties of covariance and correlation matrices of $\hat{\gamma}$ estimated using the delta method are examined. For this section, the population covariance matrix $\mathbf{V}[\hat{\gamma}]$ was estimated without using the delta method, but rather by numerical estimation from 10,000 samples of $\hat{\gamma}$ (each originating as an independent random sample of size c from $\mathbf{s}(T)$, with a specified λ). $\mathbf{C}[\hat{\gamma}]$ was then calculated from $\mathbf{V}[\hat{\gamma}]$. As noted earlier $\mathbf{V}'[\hat{\gamma}]$ is an approximation of the population covariance matrix of $\hat{\gamma}$, $\mathbf{V}[\hat{\gamma}]$ the population covariance matrix of $\hat{\gamma}$. Let $\hat{\mathbf{V}}'[\hat{\gamma}]$ be an estimate of the population covariance matrix calculated using the delta method, but replacing $\mathbf{s}(T)$ with $\hat{\mathbf{s}}$. Here we examine the bias of entries in $\mathbf{V}'[\hat{\gamma}]$, $\mathbf{C}'[\hat{\gamma}]$, $\hat{\mathbf{V}}'[\hat{\gamma}]$ and $\hat{\mathbf{C}}'[\hat{\gamma}]$ for combinations of c and λ .

4.4.1 Bias in entries in $\mathbf{V}'[\hat{\gamma}]$ and $\mathbf{C}'[\hat{\gamma}]$, estimated with $\mathbf{s}(T)$.

To examine the extent and pattern of bias introduced by the delta method when calculating the population covariance matrix, we plotted the elements of $\mathbf{V}[\hat{\gamma}]$ against $\mathbf{V}'[\hat{\gamma}]$. With c = 1000 all these plots were linear with a gradient indistinguishable from one. This was also the case with $\lambda_{1/4}$ and c = 100. Consequentially correlation matrices, $\mathbf{C}'[\hat{\gamma}]$, estimated from $\mathbf{V}'[\hat{\gamma}]$ in these instances were also indistinguishable from $\mathbf{C}[\hat{\gamma}]$. These results indicate that the delta method had converged to give accurate estimates of $\mathbf{V}[\hat{\gamma}]$. The gradient of the plots with c = 100 and $\lambda_{1/2}$ then λ_1 were 1.1 and 1.15 respectively, indicating underestimates of the entries in $\mathbf{V}[\hat{\gamma}]$ by approximately 10% and 15% respectively. Closer examination of the size of the underestimate

of each entry in $\mathbf{V}'[\hat{\gamma}]$ in percentage terms showed that those near zero did not all have the same percentage of bias (plots not shown). This showed up in the corresponding plots of $\mathbf{C}[\hat{\gamma}]$ vs $\mathbf{C}'[\hat{\gamma}]$ as a slight s bend to what would otherwise be a straight line constrained to go through 1, 0 and -1. Maximum departures of these correlation matrix entries with $\lambda_{1/2}$ and λ_1 (0.04 and 0.02 respectively) are considered insignificant.

4.4.2 Error and bias in $\hat{\mathbf{V}}'[\hat{\gamma}]$ estimated from random samples, $\hat{\mathbf{s}}$.

We now calculate the covariance matrix of $\hat{\gamma}$ (via the delta method) repeatedly (10,000 times) from samples of $\mathbf{s}(T)$. The marginal distributions of entries in $\hat{\mathbf{V}}'[\hat{\gamma}]$ show similar features to the marginal distributions of $\hat{\gamma}$, that is, some discreteness, smoothing and skewness for small c or λ. This may be better understood by realising that the calculation of these covariance matrices can also be achieved by applying a Hadamard conjugation to a vector representation of $\mathbf{V}[\hat{\mathbf{s}}]$ (see appendix 4.1). However, since entries in $\hat{\mathbf{V}}[\hat{\mathbf{s}}]$ are products of multinomial cells, they are more smoothed and skewed than entries in the samples themselves ($\hat{\mathbf{s}}$). The increased smoothness and skewness of entries in $\hat{\mathbf{V}}'[\hat{\gamma}]$ vs $\hat{\gamma}$ is also due to the covariance matrix of ρ being a quadratic function of the covariance matrix of $\mathbf{r}$ (which can show more skewness than the logarithmic function).

In contrast to the underestimation of entries in $\mathbf{V}[\hat{\gamma}]$ by $\mathbf{V}'[\hat{\gamma}]$, $\hat{\mathbf{V}}'[\hat{\gamma}]$ appears to be considerably closer to an unbiased estimator of $\mathbf{V}[\hat{\gamma}]$. Whereas entries in $\mathbf{V}'[\hat{\gamma}]$ are biased downwards in absolute magnitude, $\hat{\mathbf{V}}'[\hat{\gamma}]$ tends to give slight overestimates of the absolute magnitude of entries in $\mathbf{V}[\hat{\gamma}]$. This is understandable because application of the delta method, which involves multiplying pairs of estimated gradients, shows positive bias when applied to gradient estimates which are already skewed by the natural log function. In effect, these two sources of bias cancel each other to a large degree, to reduce the overall bias.

The sampling variance of $\hat{\mathbf{V}}[\hat{\mathbf{s}}]$ is a larger contributor to the mean square error (MSE) of entries in $\hat{\mathbf{V}}'[\hat{\gamma}]$ than is the bias. In all cases, including the most extreme with $c = 100$ and λ_1, the bias contributed to less than 1.5% of the MSE of any entry in $\hat{\mathbf{V}}'[\hat{\gamma}]$. We could discern no correlation between the bias (as a proportion of the MSE) with either the magnitude of entries in $\mathbf{V}[\hat{\gamma}]$ or the observed variance of entries in $\hat{\mathbf{V}}'[\hat{\gamma}]$ (plots not shown). Thus the influence of bias appears to be quite evenly distributed throughout all entries in $\hat{\mathbf{V}}'[\hat{\gamma}]$.

Converting variances to standard deviations (a square root transformation) removes the skewness in the marginals of these covariance matrix entries except when $c = 100$. In this case, there remains skewness due to the natural logarithm transformation applied in going from $\mathbf{r}$ to ρ. Entries in $\hat{\mathbf{C}}'[\hat{\gamma}]$ are generally closer to normal in their marginal distributions than entries in $\hat{\mathbf{V}}'[\hat{\gamma}]$ (except when distorted by the bounds of 1 or -1), yet still have large variances. For example, the correlation entry of γ_5 and γ_6 in $\hat{\mathbf{C}}'[\hat{\gamma}]$ had 2.5% and 97.5% quantiles of approximately -0.5 to 0.6 and mean 0.05 when estimated with $c = 100$ (and $\lambda = 1$).

In a given sample, the variance entries in that sample's covariance matrix $\hat{\mathbf{V}}'[\hat{\gamma}]$ (that is the diagonal) showed a high degree of correlation (all > 0.85) with corresponding entries in $\hat{\mathbf{s}}$. This indicates that the raw count for a character bipartition is the dominant variable in determining the variance of the same bipartition in $\hat{\gamma}$. This is not unexpected given the relatively small correlations between entries within $\hat{\gamma}$. As $\lambda \rightarrow 0$, then variance of $\hat{\gamma}_i$ tends to the variance of $\hat{s}_i$. The correlation of the i-th variance in $\hat{\mathbf{V}}'[\hat{\gamma}]$ with $\hat{\gamma}_i$ in these simulations was always distinctly less than that with $\hat{s}_i$. Again this is expected since $\hat{\gamma}$ aims to represent the expected values of bipartitions after taking into account multiple changes, but not their variances, whereas $\hat{s}_i$ by its multinomial nature reflects both in its magnitude. Most entries in $\hat{\mathbf{V}}'[\hat{\gamma}]$ also show high correlations with many other entries in the same matrix. This is understandable since the order $(2^{t-1})^2$ entries in the covariance matrix, are being estimated from just 2^{t-1} values in $\hat{\mathbf{s}}$. The high correlation between the size of $\hat{s}_i$, and the variance of $\hat{\gamma}_i$ suggests that the variance of $\hat{s}_i$ may be a useful approximation to the variance of $\hat{\gamma}_i$ when $\hat{\mathbf{V}}'[\hat{\gamma}]$ is prohibitively large to calculate and rates of change are low.

4.5 ESTIMATING $\mathbf{V}'[\hat{\gamma}]$ WHEN COMPENSATING FOR UNEQUAL RATES ACROSS SITES

In chapter 2 we derived formulae that can be used to make the transformation from $\mathbf{r}$ to ρ when rates of substitution vary across sites (see table 2.2 especially). All the steps in calculating $\mathbf{V}'[\hat{\gamma}]$ are the same as those for the i.r. model, except that when the delta method is applied. So to calculate $\mathbf{V}'[\hat{\rho}]$ when using unequal rates across sites (URAS) correction formulae, the terms $1/r_i$ and $1/r_j$ in formulae 2.9 and 2.10 are replaced with the appropriate first derivatives of the correction formulae at r_i and r_j. As already discussed in chapter 2, with continuous distributions of rates across sites, the i.r. model is a limiting form as the coefficient of variation (c.v.) of the distribution of rates across sites goes to zero. As the c.v. becomes larger to give unequal rates across sites, then the non-linearity of the appropriate transformation becomes more pronounced. Consequently, features like bias and skewness in $\hat{\gamma}$ will be exacerbated relative to the i.r. model (for a given sequence length, and weighted tree combination), yet the general features of the marginal distributions of $\hat{\gamma}$ will be like those described already in section 4.3.

4.5.1 First derivatives of closed form URAS correction formulae

The maximum likelihood formulae for transforming $\hat{\mathbf{r}}$ to $\hat{\rho}$ when sites follow a gamma (Γ) distribution is, from table 2.2, $\hat{\rho}_i = k(1 - \hat{r}_i^{-1/k})$ (where k is the shape parameter, taking any positive value). The first derivative of this formulae with respect to r_i is,

$$\rho'_i = -k \times -(1/k)r_i^{-(1/k)-1} = r_i^{-(1/k)-1} \qquad (4.5.1\text{-}1)$$

Using the delta method, our estimate of the variance of $\hat{\rho}_i$ is thus $(r_i^{-(1/k)-1})^2 Var[\hat{r}_i]$ which equals $(r_i^{-(2/k)-2})Var[\hat{r}_i]$. As $k \rightarrow \infty$, this Γ distribution transformation converges to the natural log transform, and accordingly $V'[\hat{\rho}]_{ii} \rightarrow (r_i^{-2})Var[\hat{r}_i]$ (the delta method variance of the i.r. model, equation 4.2.4-1). Correspondingly the covariances of entries in $\hat{\rho}$ (by the delta method) are,

$$V'[\hat{\rho}]_{ij} = (r_i^{-(1/k)-1})(r_j^{-(1/k)-1})\, V[\hat{r}]_{ij},$$

which correspondingly converges to equation 4.2.4-2 as $k \rightarrow \infty$.

The correction formulae, assuming an inverse Gaussian distribution of rates across sites, is from table 2.2, $\rho_i = 0.5d[1-\{1-(\ln(r_i)/d)\}^2]$. Expanding out this expression we have,

$$\rho_i = 0.5d[1-\{1- 2(\ln(r_i)/d + (\ln(r_i)/d)^2\}] = 0.5d[- 2(\ln(r_i)/d + (\ln(r_i)/d)^2]$$

Differentiating gives,

$$\begin{aligned}\rho'_i &= 0.5d[- 2/(r_i d) + 1/(r_i d).2(\ln(r_i)/d)] \\ &= 1/r_i - 1/r_i.\ln(r_i)/d \\ &= 1/r_i[1-\ln(r_i)/d]. \qquad (4.5.1\text{-}2)\end{aligned}$$

Again as $d \rightarrow \infty$, derivative 4.5.1-2 $\rightarrow 1/r_i$, since the underlying distribution of rates across sites then converges to the i.r. model.

Next, we derive the variances and covariances of $\hat{\rho}$ when assuming a proportion of invariant sites. Assuming that the distribution of rates across sites has variable sites, x, plus a fixed proportion of invariant sites, p_{inv}, then the correction formula is $M_x^{-1}[(r_i - p_{inv}) /(1- p_{inv})]$ (eq 2.3.4-6). Thus, using the delta method, the variance of $\hat{\rho}_i$ for a model with a proportion (p_{inv}) of invariant sites and all other sites i.r. is:

$$\begin{aligned}Var'[\hat{\rho}_i] &= ((r_i- p_{inv})/(1- p_{inv}))^{-2} Var[\hat{r}_{i\text{-}pinv}] \\ &\approx ((r_i- p_{inv})/(1- p_{inv}))^{-2} Var[\hat{r}_i].1/(1- p_{inv})^2 \qquad (4.5.1\text{-}3)\end{aligned}$$

while for the covariance:

$$\begin{aligned}Cov'[\hat{\rho}_i, \hat{\rho}_j] &= ((r_i- p_{inv})/(1- p_{inv}))^{-1}((r_j- p_{inv})/(1- p_{inv}))^{-1} Cov[\hat{r}_{i\text{-}pinv}, \hat{r}_{i\text{-}pinv}] \\ &\approx ((r_i- p_{inv})/(1- p_{inv}))^{-1}(r_j- p_{inv})/(1- p_{inv}))^{-1} Cov[\hat{r}_i, \hat{r}_j].1/(1- p_{inv})^2 \qquad (4.5.1\text{-}4)\end{aligned}$$

where $r_{i\text{-}p_{inv}}$ is the value of r_i when removing the effect of the invariant sites, and the factor $1/(1- p_{inv})^2$ approximately accounts for the conversion of r_i into $r_{i\text{-}pinv}$ by $r_{i\text{-}pinv} = (r_{i\text{-}pinv})/(1-p_{inv})$. The approximation $Var[\hat{r}_i](1/[1- p_{inv}]^2)$, gives a slight overestimate of $Var[\hat{r}_{i\text{-}pinv}]$, but results in

standard deviations within 10% of their true value for $p_{inv} \approx < 0.4$ and $r_i \approx > 0.2$. The variance of just the variable sites estimating r_i, $Var[\hat{r}_{i\text{-}pinv}]$, can be estimated exactly as, $([1-r_{i\text{-}pinv}]/2)(1-(1-r_{i\text{-}pinv}]/2) / c(1-p_{inv})$, where $r_{i\text{-}pinv} = (r_i-p_{inv})/(1-p_{inv})$ (this is because the marginal distribution of r_i is a binomial distribution, a described in section 4.23). Unfortunately, there is no such convenient formula for the covariances of just the variable sites, given the variances of the mixture of variable and invariant sites, since these are an uneven non-linear function of the observed sequence pattern proportions. (That is, when a proportion of constant sites are removed, the estimated covariances of $\hat{s}_i$, and $\hat{s}_j$ change as,

$$\hat{s}_i.\hat{s}_j/c \rightarrow \hat{s}_i/(1- p_{inv}) \times \hat{s}_j/(1-p_{inv}) / c/(1-p_{inv}) = \hat{s}_i.\hat{s}_j/(c(1-p_{inv})^3),$$

while the variances change as,

$$\hat{s}_i(1 - \hat{s}_i)/c \rightarrow \hat{s}_i(1\text{-pinv}) \times(1 - \hat{s}_i /(1-p_{inv})) / (c/(1-p_{inv})) = \hat{s}_i(1- \hat{s}_i/(1\text{-pinv}))/(c(1-p_{inv})^2),$$

so the change is not uniform across all entries in our estimate of $\mathbf{V}[\hat{\mathbf{s}}]$). Fortunately, it is usually not prohibitive to recalculate $\mathbf{V}[\hat{\mathbf{s}}_{i\text{-}pinv}]$ after removing a proportion, p_{inv}, from $\hat{s}_0$, multiplying all $\hat{s}_i$ by $1/(1-p_{inv})$, recalculating c as $(1-p_{inv})$, then using equation 4.2.3-1 to estimate $\mathbf{V}[\hat{\mathbf{r}}_{i\text{-}pinv}]$. More generally, this same approach needs to be used to estimate the variance of any mixture of invariant sites plus variable sites, with the variable sites following a distribution with moment generating function M. That is,

$$Var'[\hat{\rho}_i] = (1/(M^{-1})')^2 Var[\hat{r}_{i\text{-}pinv}] , \qquad (4.5.1\text{-}4)$$

where $(M^{-1})'$ is the first derivative of the path correction formula with r_i replaced by,

$$r_{i\text{-}pinv} = (r_i - p_{inv})/(1- p_{inv}).$$

4.5.2 How unequal rates across sites affect accurate distance estimation

Figure 4.6 is an illustration of how much unequal rates across sites increase not only the corrected pathset length, but also its variance (or standard deviation). The parameter values shown for the Γ distribution correction, and for the invariant sites correction, are practically identical to those estimated from the conserved regions of anciently diverged rRNA (see chapters 2 and 3). Only the variance for $r_i > 0.3$ is shown in figure 4.6a, since for lesser values of similarity it becomes very large. The variance of the uncorrected data (marked *u* in fig. 4.6a) is the same as the variance of $\hat{r}_i$, which, as already mentioned, has a scaled binomial distribution.

It is crucial to consider the size of relative errors, after making non-linear transformations. Figure 4.6b shows the relative accuracy of pathlength correction (the ratio of ρ_i to it's estimators s.d.) with respect to the similarity, r_i (shown here for a sequence length of 1000 sites). This measure of relative error, which is not discussed in phylogenetics, can be considered a type of 'signal-to-noise' ratio. This type of measure should be given more prominence. It is a much more useful predictor of the performance of transformed distance tree estimation methods, than quoting raw standard errors. Just because one transformed value has a larger variance than another, it need not be a worse estimator if its mean value has also increased (just as we see in figure 4.6).

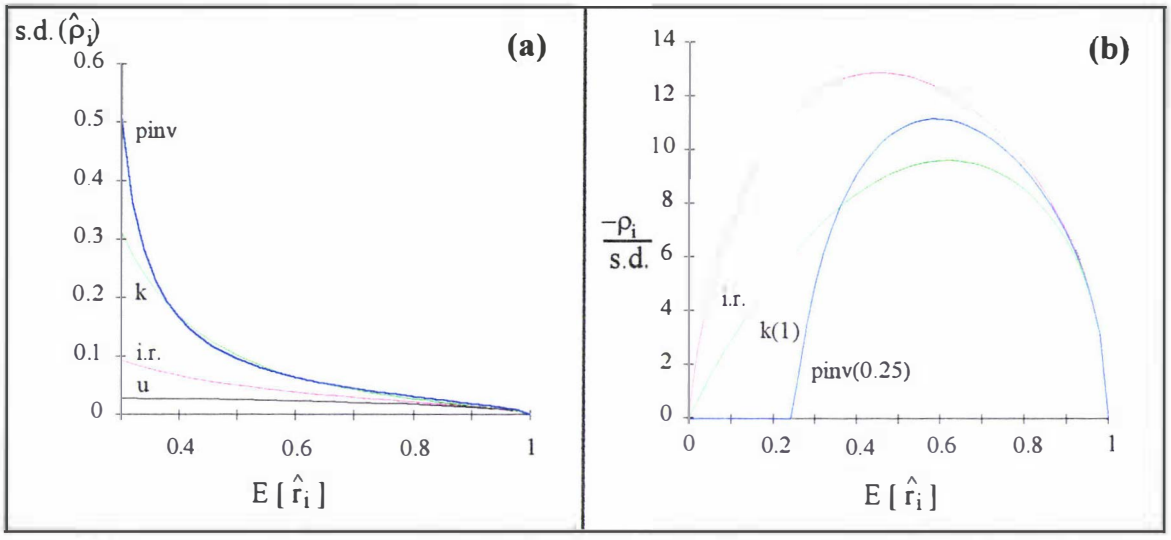


FIGURE **4.6a** The standard deviation of corrected pathset lengths with respect to the expected value of the observed dissimilarity, $\hat{r}_i$ (standard deviations shown are for a nominal sequence length of 1, and evolutionary distances are becoming larger moving to the right, or smaller r_i values). The symbols are p_{inv} for proportion 0.25 of sites invariant (blue lines), *k* for an underlying gamma distribution shape $k = 1$ (green), *i.r.* the identical rates correction (red), and *u* uncorrected (black). The lines for pinv include resetting, *c* to $c(1-p_{inv})$. **4.6b** A graph of $-\rho_i$ divided by the standard deviation of $\hat{\rho}_i$, plotted against the expected value of $\hat{r}_i$ (for sequence of length 1000). Here the higher the y-entry, the more accurately, in terms of parentage error, $\hat{\rho}_i$ values can be predicted (given just sampling error).

Using a signal-to-noise measure is important to obtain a fair idea of how different models (here distinct underlying distributions of rates across sites) will behave. Figure 2.6a shows that the optimal accuracy with which a pathset-length can be predicted occurs when r_i is approximately 0.55 (for the 2-state Poisson i.r. model), which translates to an observed distance of 0.275. At this value, with a sequence of length 1000, the average stochastic error is expected to be approximately 1/12.7 (or 8%) of the distance measured (not stunningly accurate, but not awful either). For models with a proportion of invariant sites, the accuracy with which a distance can be measured is always worse than that under the i.r. model. In contrast, while a Γ distribution of rates across sites always results in lower accuracy than the i.r. distribution, it can be both better and worse than a proportion of invariant sites (worse for the short distances, but becoming better on the largest distances).

The results in figure 4.6b can be used to predict the best accuracy for measuring distances, which should will also coincide largely with not only the best tree estimation, but also the most accurate inference of relative branching times (assuming, of course, that the model holds). Any distribution of rates across sites will decrease the maximum accuracy from the optima achievable under an i.r. model as will any deviation from equifrequency (even if stationarity holds). Put another way, if you were looking for the ideal sequences with which to analyse a particular phylogenetic problem, a rough guide is that you would want sequences:

(1) with very little site to site variation in rates,

(2) near equal frequencies of the 4 bases,

(3) observed distances neither too big nor too small (and hopefully mostly falling around 0.27 with 2-state data, then higher the more states your sequences have).

This, of course, is only a rough guide since tree estimation algorithms combine distances in many different ways. As shown later, it is sometimes the smaller or larger distances (not just the average distance) that you want measured most accurately but none-the-less, this approach conveys a useful sense of what is going on.

The issue of the relative accuracies expected under an invariant sites / i.r. model, and a Γ distributed rates model, is interesting. In figure 4.6b, notice how the accuracy of the invariant sites / i.r. transform (when applicable) is generally higher than that of the Γ transformation. Even increasing the proportion of invariant sites to 0.3 to make this distance more additive when the true underlying distribution of rates is Γ with $k = 1$, the accuracy remains very similar than that of the Γ transformation for short to moderate distances. It only deteriorates at the larger distance estimates (in this case $r_i = 0.35$, $d_{obs} = 0.325$, which gives $d_{obs\text{-}p_{inv}} = 0.433$, or a transformed distance of > 1.00). This suggests that an invariant sites-distance transform will, with long sequences, perform quite comparably to corresponding Γ distribution transformation. A further interesting point is that when the invariant sites model begins to deteriorate in its relative performance, it does so quite rapidly. Soon after this deterioration, observed r_i values which will not allow the usual logarithmic transform (i.e. negative logarithmic arguments) are expected to be encountered increasingly often.

In chapter 3, the exact same transformations illustrated here were used on the transversional changes in the 16S-like rRNA sequences representing the tree of life (these were done by considering the transversional distances under the 4-state Kimura 2 or 3 ST model, see appendix 2.6). According to predictions, the bootstrap support on neighbor joining trees was very similar, even after removing up to 30% of the constant sites. In this case, no negative logarithm values were encountered in the original data, and they were also very rare in the bootstrap samples. This is in agreement with the observation that the longest paths through the tree built from transversional changes alone tended to be less than 0.5 (i.e. the transformed distance) even with 30% of constant sites removed.

The situation was slightly different with the 4-state invariant sites-LogDet transform applied to this data (see section 3.7.3 and 3.7.4). While no negative logarithmic arguments were encountered upon application to the observed data, they became quite frequent when more than 30% of constant sites removed (occurring one or more times in 1/10 of bootstrap replicates with all constant sites removed). This suggests that the invariant sites-LogDet transform would probably have had similar sampling variance to a transform taking into account the exact distribution of rates across sites (if such a transform existed), but then deteriorated with approximately 30% or more of the constant sites removed. This is compatible with what is seen in figure 3.11. This agreement in turn, gives further confidence that the evaluations were reasonable, but with the invariant sites-LogDet there can be a rather sudden drop in support past a critical number of constant sites removed. In this region, the invariant sites LogDet transform

may not reflect the continued higher bootstrap support of other, sometimes equally valid, procedures for tree estimation (including the general invariant sites / i.i.d. ML model). Thus, as a rule of thumb, it may be appropriate to take the best estimate of support in such situations at the point when negative logarithmic arguments are just becoming noticeable (say 1/50) in the bootstrap replicates. Of course, the more shallow branching lineages are not so affected by this factor, so support for them can still be accurately gauged with further constant sites removed.

It is useful to consider other ways of comparing distance signal-to-noise ratios. In figure 4.6 the accuracy of distance estimation was plotted against r_i (a similarity measure). Many tree estimation algorithms, however, compare the standard deviations of the corrected distance directly with its size (e.g. weighted least squares tree building procedures, see Felsenstein 1982, 1988). In this case a plot of signal-to-noise ratio verses the true pathset length, ρ_i, should reveal further features of the interaction of distance estimation and tree selection. This is shown in figure 4.7a and several points become apparent. The invariant sites model has errors which are proportionately worse than those under the i.r. model. This is expected since the variable sites are i.r., but constitute only $1\text{-}p_{inv}$ of the sequence. Accordingly, the variance is higher by a factor of $1/(1\text{-}p_{inv})$, which in this case is 1/0.75 or 1.33. Thus, the standard deviation (s.d.) is higher by $\sqrt{1.33} \approx 1.15$, so the signal-to-noise ratio is worse by 1-1/1.15 = 1-0.87 or 13% lower. It is obvious, also, how the peak accuracy is achieved at a moderately large true distance of $\delta \approx 0.275$ (or $r_i \approx 0.45$) with the i.r. model, or invariant sites model, and it does not fall off quickly; something not so immediately evident in figure 4.6b.

Figure 4.7a also shows an important property of the gamma distribution and other continuous, tailed distributions (e.g. the lognormal or the inverse Gaussian). Under these models the relative accuracy of path length corrections remains much more constant between long and short paths when rates across sites become highly spread out. This effect is clearly seen as the difference in the curves for $k = \infty$ (the i.r. model, c.v. = $1 / k^{0.5} = 0$) vs $k = 1$ (c.v. = 1) vs $k = 0.25$ (c.v. = 2). This trend is predictable; as time goes by, sites with progressively slower rates become more important to inferring the true distance, and with these skewed uni-modal distributions there can be a good supply of such slowly evolving sites, even at relatively large distances. In the authors experience, fitting a model with an underlying gamma distribution of rates across sites to DNA sequences with functional constants (near neutral sites removed) typically returns an estimate of k between 1 and 0.25 (see also Uzzel and Corbin 1971, Wakeley 1993, Yang 1993, Waddell and Penny 1995 for similar findings). Accordingly, if real sequences have this sort of distribution of rates across sites, then it can be expected when using distances estimators which assume this type of model, that the signal-to-noise ratio will be relatively flat for transformed distances between 0.1 and 1.5 or more. This feature is even more pronounced with 4 or more states.

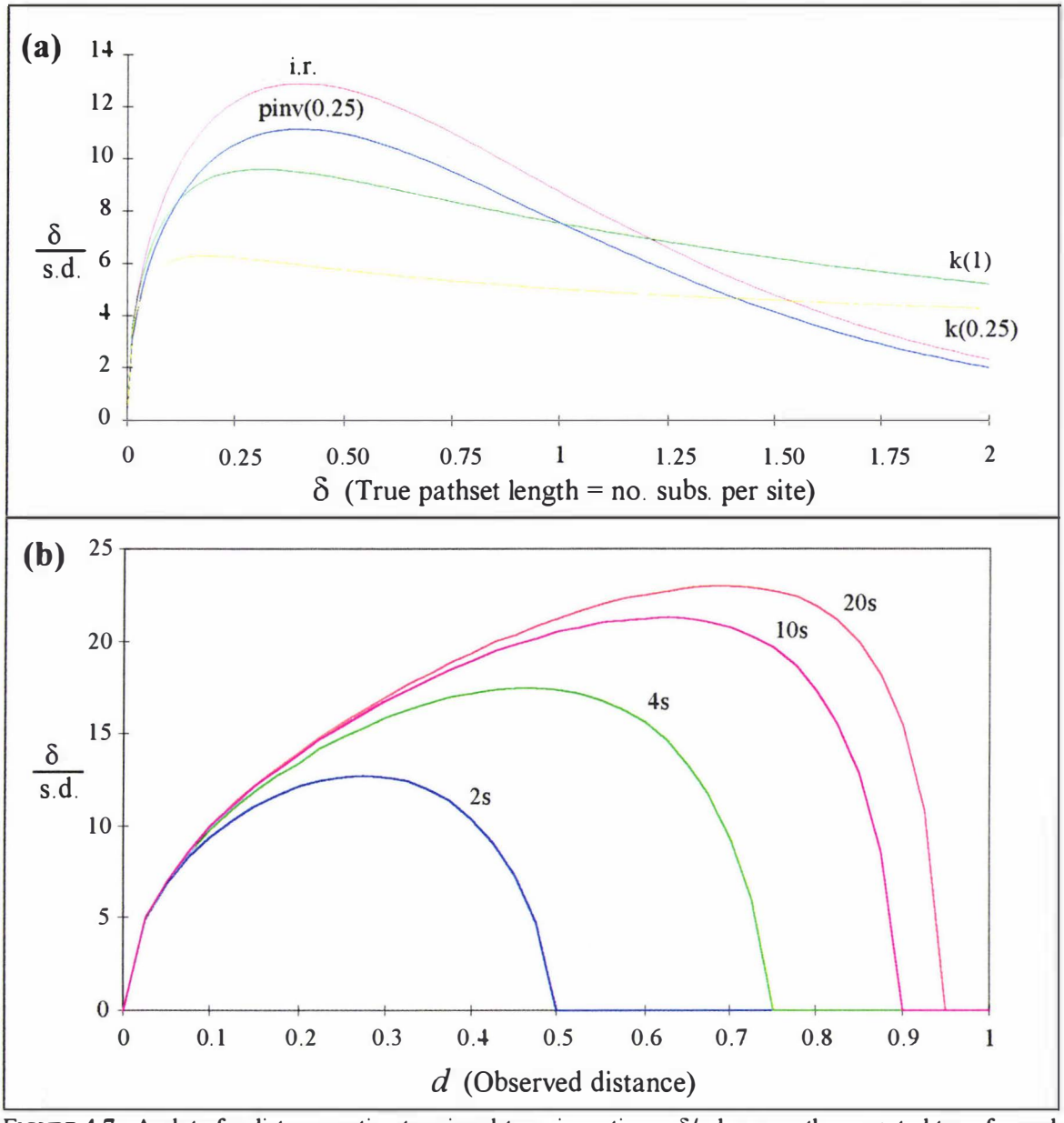


FIGURE **4.7a** A plot of a distance estimates signal-to-noise ratio, or δ/s.d, verses the expected transformed distance, δ (measured in substitutions per site), under 2-state Poisson models. The identification of the models is given in figure 4.6. **4.7b** The signal-to-noise ratio for Poisson distances when the data are 2-state (blue), 4-state (green), 10-state (purple), and 20-state (red), plotted against the observed distance.

Importantly, the Fitch-Margoliash fit criterion (Fitch and Margoliash 1967, Felsenstein 1982, 1988, 1993) is a weighted least squares method that makes the assumption that the s.d. is in a constant ratio to the true distance. This assumption of that criterion is not always noted, as it is usually expressed as minimising the sum of {transformed observed distance - tree distance}^{2}/tree distance. Consequently, it may be a reasonable criterion to get some indication of which distribution of rates across sites give the best expected additivity, taking account of the increased sampling variances that the more extreme transformations impart. Hopefully it should allow the Γ distributions shape parameter k (when less than < 1) to be optimised moderately well by the Fitch-Margoliash criterion, especially if the correlations between distances do not change appreciably as k changes (something which needs to be verified). This observation is of practical use since the only generally available weighted least squares program appears to be the "Fitch"

in the PHYLIP package (Felsenstein 1993). More generally, this program also allows fitting of distances to trees by minimising the sum of (estimated distance - tree distance)^2 / (tree distance)^y, where y is specified by the user. By choosing y appropriately, it should also be possible to refine the assumed relationship between random errors and size of distance. This fine tuning of the algorithm should offer more reliability in tree selection, while the algorithm itself involves less computational cost than having to store and recalculate many independent distance variance estimates. Up to the peak of each curve in figure 4.7a the variance is increasing more slowly than the square of the distance (else the first part of the curve would be perfectly linear rather than concave), and only begins to increase in proportion to the square of the distance once the peak is reached. For these shorter distances, the optimal value of weighting factor y will fall between 0 and 2 (these values are respectively the Cavalli-Sforza and Edwards 1967 least squares weighting which assumes constant variance, and that of Fitch and Margoliash 1967). A factor of about 1 would seem a reasonable compromise value, if all distances fell between zero and the peak signal-to-noise distance. Beyond the peak of distance / s.d. the variable y needs to rise above 2 in order to give a distance weighting in line with the expected increase of variances with distance.

It is also important to note that the signal-to-noise ratio improves substantially when the data has more states, as is clearly shown in figure 4.7b. The trend is for the peak signal-to-noise ratio to become higher with the number of states, and to then occur at larger distances (some times surprisingly large distances). Here is a list of the points at which these maxima occur, and also the largest and smallest distances at which the signal-to-noise ratio is at least 50% as large as at the peak value.

	max. s to n ratio	first hits 50%		reaches peak		drops under 50%	
		observed d	(corrected)	observed d	(corrected)	observed d	(corrected)
2-state	12.7	0.043	0.045	0.275	0.399	0.460	1.263
4-state	17.5	0.079	0.083	0.463	0.720	0.705	2.110
10-state	21.3	0.116	0.125	0.620	1.051	0.864	2.899
20-state	23.0	0.134	0.144	0.692	1.238	0.921	3.315

So, under the 4-state Poisson model, the maximum signal-to-noise ratio is estimated to be 17.5 with c = 1000. The signal-to-noise ratio first hits 50% of this maximum when the observed distance reaches 0.079 (which corresponds to a corrected distance of 0.083 under this model); it peaks at δ = 0.720 and does not drop more than 50% until a fairly huge distance of 2.1. The figures for the 20-state model, which may be considered a very simple amino acid model (e.g. see Cao *et al.* 1994), are even more stunning. That the number of states gives models this characteristic shows up an important feature not usually considered: it is substitutions back to the original state, and not multiple substitutions at a site which are the major problem. In this way distances are behaving very like character based methods such as parsimony or maximum-likelihood, where multi-state characters are generally fine, and it is parallelisms and convergences which are the major bugbear.

The 10-state model, shown in figure 4.7b, may be considered to better approximate protein sequences, where there is unequal amino acid usage frequencies. That is, a *b* value of about 9/10, rather than 19/20, is usually more appropriate for an equal-input model formula such as $\delta = -b.\ln(1-d_{obs}/b)$ (e.g. Nei 1987). Of course the same thing can be said of the 4-state model when unequal nucleotide frequencies are encountered (e.g. for the Horai *et al.* 1992 data of section 1.9.2, *b* estimated for the variable sites would be $1 - \Sigma f(i)^2 = 1 - (0.251^2 + 0.385^2 + 0.089^2 + 0.275^2 = 0.705$ rather than 0.75). Of course, quite unequal transition probabilities between states will also lower the maximum signal-to-noise ratio. A highly unequal transition to transversion rate is one example. Such factors act approximately as if there were fewer states, and unless the effect of increased parallelisms and convergences is countered (for instance, by down weighting the most frequent changes, e.g. Schöniger and von Haeseler 1993, and appendix 2.6) then the peak is not so high, and the signal-to-noise ratio decays more quickly than in the Poisson case. In such a situation where transitional changes are all but saturated, the signal-to-noise ratio will be higher at large distances under a 2-state purine / pyrimidine model, than the full 4-state model (the same thing will ultimately hold for sequence based ML tree selection, although we suspect at slightly larger distances due to its greater stability to sampling error).

An important point is that the trends shown in figure 4.7b do not automatically mean it is better to keep coding data as amino acid changes. The sequences being compared are not identical, i.e. the 4-state sequences are not taken from the amino acid codons (of course if there were we would be comparing 3000 nucleotide sites with 1000 codons). However, the graphical method of analysis developed here can be extended to study the question of when distances are best estimated by nucleotide sequences verses amino acids, or their 61 codons (excluding the 3 stop codons). It is easy to see a simple approximation under these Poisson models. Our model is one in which all changes in the sequences are Poisson, all amino acids have 3 codons, there are 4 stop codons, and all three coding positions evolve at the same rate (thus changes at codon positions are almost independent, but for the avoidance of the stop codons). Here, it is approximately correct to compare using 4-state Poisson sequences which are 3,000 base pairs long, with amino acid sequences (under the 20-state Poisson model) which are 1,000 codons long. This will make the signal-to-noise ratio for the 4-state sequences the square root of three (or about 1.73) times as large as they are in figure 4.7b. Thus, at short observed distances such as 0.134, the signal-to-noise (S/N) ratio using the nucleotide sequences will be approximately 19.5, while that for the amino acid sequence will be 23.0 / 2 = 11.5. The peak for the amino acid sequences will remain unchanged from the list given earlier (the table on the previous page), that is 23.0, while that for the nucleotide sequences will climb to 1.72×17.5 or 30.1. However at the larger distances, the S/N ratio of the amino acids sequences will catch, then steadily exceed that of the nucleotide sequences. Having unequal rates at the three codon sites, will tend to exaggerate these differences. By an extension of this approach it should be possible to make predictions of the performance of these two distinct treatments of the data, without needing to resort to time consuming and less theoretically grounded simulations (although simulations may be important to check for the effects of bias etc. with short sequence lengths).

4.5.3 Knowing the model, we can estimate even very large distances accurately

Here we look in more detail at the implications of signal-to-noise ratios which remain good up to substantial distances, often in excess of an average of 1.0 substitutions per site. As figure 4.7a and b show that the signal-to-noise ratio remains reasonable (relative to the peak accuracy) for distances of up to about 1.25 substitutions per site, even under the i.r. model. Under Γ distributed models, larger true distances (but smaller observed distances) retain even more favourable accuracies compared to the shorter distances. This aspect is even more marked when 4-state models are studied, with the maximum signal-to-noise ratio jumping from 12.7 to approximately 17.5 for the same length of sequence (c = 1,000) under the i.r. Jukes-Cantor model (see figure 4.8). Contrary to these findings, a number of authors have been dismissive of using these formulae to estimate distances when there has been more than one substitution per site (e.g. Jin and Nei 1990, Rodríguez *et al* 1990), with the standard reason being increasingly large standard errors (not systematic errors). This objection obviously need not hold with sequences of reasonable length (more than 1000 sites, possibly of the order of 5,000 + sites to nearly completely remove the effects of sample size bias).

The two factors which can make distance estimation deteriorate more quickly than suggested here are bias and systematic error. The other factor which our graphs do not show (being based on the delta approximation) is bias, yet bias itself falls away rapidly (it is a second order effect) beyond some critical sequence length (simulations in section 4.4 give an example of how bias disappears as sequences get longer). The more serious concern is systematic error between substantially diverged sequences; it is necessary to be suspicious of any model applied to the data, due the potential for serious departures between reality and the methods assumptions (systematic error). Accordingly, we conclude that if systematic errors remain acceptable, it is possible to estimate large distances (1 to 2 substitutions per site) as reliably as many shorter distances, given long sequences.

One situation where it may be possible to have confidence in tree building using sequences where there has been more than one substitution per site (and perhaps up to 2.5 to 3 in the largest pairwise distances) are in non-coding DNA regions. Here, the three necessary requirements of: (1) long sequences (10 kb or more). (2) distances based on a model which well approximates reality. (3) a substitution process which is uniform and not too extreme, might be satisfied. The first requirement should be easily meet by the rapid advances in sequencing technology. The second requirement needs consideration. Firstly, while intrinsic substitution rates do vary between sites, overall the coefficient of variation is probably low. For example, Waddell and Penny (1995, see also chapter 5) used ML to fit models allowing sites to follow a Γ distribution, to two stretches, each approximately 10kb in length, of non-coding nuclear primate sequence, and in each case found an optimal k value of about 4 (with no significant difference from $k = \infty$). Secondly, sites often evolve pretty much independently. Studies have found low, but measurable, correlations between sites 1 to 3 positions distant, yet hopefully such local correlations will not be a major problem with 10kb sequences (e.g. Bernstein's theorem, Rényi 1970, p. 379). The third requirement is that base compositional shifts are fairly uniform across

the sequences (e.g. all sites analysed together are from the same isocore). If this is met, then it is reasonable to hope the LogDet transform (combined with the removal of constant sites, if necessary) will be reliable. The final qualification is that base compositional shifts (and the underlying rate matrix) do not become too extreme. If this occurs, stochastic error under the model can double (or worse), and much longer sequences would be necessary to ensure an equivalent level of confidence at the largest distances.

Thus we expect that for long sequences the sampling error of LogDet distances will be very reasonable as long as the process of substitution is not extreme (e.g. all off diagonal entries of the **F** matrix approximately similar, with the diagonals also being of similar size). Examining the form of the variance estimate equation 3.5.1-1, it becomes evident that the variance will increase as entries in **F** become more uneven in size. This includes increasing asymmetry of **F**, and also more unequal base composition (note that while equation 3.5.1-1 ignores bias, this is a second order effect which decreases rapidly with sequence length; see section 3.10.2 for another examination of this issue).

A useful new finding in regard to estimating large distances, is that the variance of the LogDet transform (equation 3.5.1-1) converges to that of the Kimura 3ST distance (or to any submodel the data was generated according to) with increasing sequence length. We noticed this in our numerical studies. The equivalence of the delta method variances of the Kimura 3P (and its submodels, including the Jukes-Cantor Poisson model) with that of the LogDet, was noticed preparing graphs like that in figure 2.9. It has been verified by trying many different rate matrix forms and checking the algebraic identity under these simpler models. The error in this equivalence is expected to be of order $1/c^2$. Simulations have tended to confirm this feature. Interestingly, the invariant sites-LogDet transform may offer a slight improvement in sampling error over, for example a Γ model, when these two methods return similar transformed distances of small to moderately large size (e.g. up to 2.0 or so with two state data). This all helps to better understand the likely behaviour of a very useful transformation, which can be used with a wide variety of types of data.

To summarise, these results suggest that long homogeneous non-coding (hence typically nuclear) sequences corrected with the LogDet transform may provide reliable data for tree inference even when there is an average of more than one substitution per site. This finding is contrary to conventional wisdom in the field (e.g. Jin and Nei 1990, Kumar *et al.* 1993). Good examples of unsolved problems where highly diverged nuclear sequences may be useful are the relationships within mammals, birds, and many other vertebrate and invertebrate groups which have diversified into major lineages between 60 and 200 million years ago. The main obstacle here is that substantial numbers of deletions can make alignment and data editing difficult. Using four-fold degenerate third position sites from many proteins may offer a useful alternative if this problem is too serious. Alternatively, alignment can be helped considerably if there is a good sampling of diverse sequences from each group, as these give extra information on the more likely alignments. An added bonus is that because of the expected regularity of substitution processes, it will be more appropriate to make pairwise comparisons of all alignable sites in two

sequences, without needing to exclude a site from all comparisons if it incurs deletions in some species (as is often advisable with coding regions). This should mean a substantial reduction in the loss of effective sequence length. Of course, in practice, we would wish to test all the assumptions made here, as we should test the assumptions of any other tree estimation technique.

A further interesting speculation from these results is that when the correction matches the model, and a sufficient number of sites are used, then the maximum rate of recovery of trees from distance data will be when average path lengths fall close to the peak signal-to-noise ratio; that is, substitution rates are neither too high nor too low for the tree to be reliably recovered. This prediction is evaluated in chapter 5 and there it is shown that the complexities of tree estimation require that the prediction be refined. An alternative conjecture, that seems more likely, is that with a reliable tree building method, the best statistical efficiency in tree recovery will occur when the longest path through the tree is longer than the distance which achieves the best signal-to-noise ratio. This relaxed prediction is expected to generally hold true given that: the tree building method is consistent, distance transformations are additive in expectation, sequences are sufficiently long, the relative edge lengths on the tree are fixed (that is the general rate of evolution which is changing), and that the tree is not a star tree with all distances equal (in which case the optima will be at the peak distance). This prediction provides a useful starting point to understanding the relationship between stochastic errors and tree building algorithms based on either distances or generalised distances (i.e. the Hadamard conjugation).

Finally in this section, we consider the use of uncorrected or observed distances. Figure 4.8 shows the signal-to-noise ratios for uncorrected distances as the purple line. As can be seen in this example, the ratios become noticeably better than that of any log transformed distance once the observed distance exceeds approximately 0.1. Clearly if non-additivity is not a major problem (something which has a lot to do with tree topology and edge weights) then the observed distances in combination with reasonable tree building method should be statistically more efficient than any transformed distance method of recovering the unweighted tree. The problem, of course, is knowing when inconsistency of tree estimation could be occurring. Also, even before inconsistency of recovering the unweighted tree, estimates of edge weights, and statistical estimates of accuracy can be severely biased.

If our conjectures and assumptions in the previous paragraphs hold true, then it should be possible to predict how divergent sequences ought to be in order to best resolve a specific phylogenetic problem (and then use this estimate to help select regions for sequencing). Further, figure 4.8, for example, suggests there is some latitude for error in this estimate as the signal-to-noise curves are fairly flat topped. Because systematic errors due to violations of the assumptions of distance transformations are a generally unknown quantity, it would be sensible to chose distances at the shorter, rather than the longer, end of any such a range of estimates (i.e. to the left in figure 4.8).

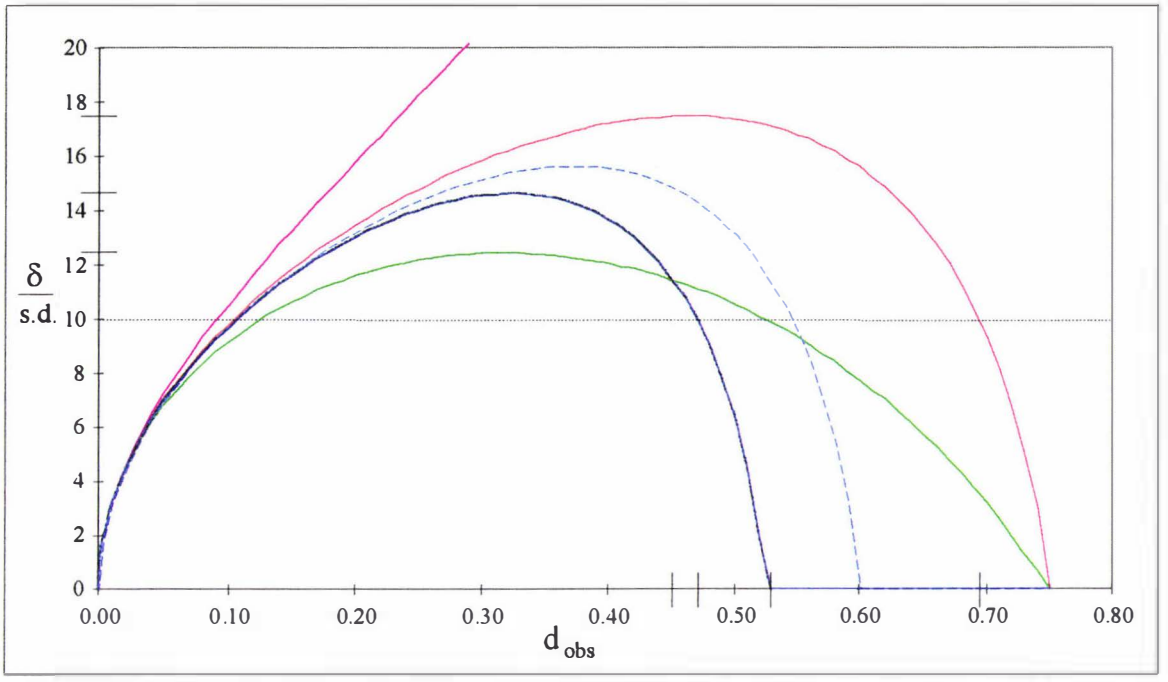


FIGURE 4.8 A plot of the observed path length, d_{obs}, versus its signal-to-noise ratio (measured as δ / s.d) measured under four different 4-state Poisson models (with a nominal sequence length of 1000). The red line is the i.r. model, blue the invariant sites model with $p_{inv} = 0.3$ (and $p_{inv} = 0.2$ for the dashed blue line marked in for visual comparison), green the gamma distribution of rates cross sites ($k = 1$), while the purple line is the signal-to-noise ratio of the uncorrected distances. An important point discovered in preparing this graph is that replacing the Jukes-Cantor (J.C.) correction formulae with the LogDet equation and its estimated variance (equation 3.5.1-1), yields an identical curve to that of the i.r. / J.C in this graph (the same also holds for the invariant sites-LogDet vs the invariant sites-J.C. transform). Considering the signal-to-noise ratio, the maximum values are 17.48 for the i.r. model, 14.62 for the $p_{inv} = 0.3$ model, and 12.44 for the gamma ($k = 1$) model. The horizontal line is at a signal-to-noise ratio of 10, which is almost exactly the signal-to-noise ratio for an observed distance of 0.10 ($\delta = 0.107$ subs. per site corrected by J.C. formula). Following this line to the right, you can see the largest observed distance at which has a signal to noise ratio better than an observed distance of 0.1. The result is: $d_{obs} = 0.695$ for the i.r. model ($\delta = 1.96$); $d_{obs} = 0.47$ with i.r. / $p_{inv} = 0.3$ ($\delta = 1.69$), and lastly $d_{obs} = 0.525$ for the gamma distribution ($\delta = 1.75$) (each δ estimated under the matching model).

4.5.4 Can data editing improve consistent tree building methods?

This section considers a fundamental issue in phylogenetic practice: Given a set of sequences, are there some variable sites I should exclude before performing an analysis? Here we approach this question from the viewpoint of statistical accuracy (or sampling error), assuming that we have a model which is consistent across the whole parameter space. That is, here we ignore issues of editing for consistency, for example excluding invariant sites, regions of atypical GC content, gaps, and other factors which will cause systematic error (see chapter 3 for examples of editing to improve consistency, while Bull *et al.* 1993 examine this question theoretically).

Evaluations of sampling error will be made on a simple illustrative model, a sequence with just two sorts of sites. Two thirds of the sites evolve at a lower rate (set *a*), while one third evolve at a higher rate (set *b*). The mean rate is fixed to one, so there is just one free parameter, namely the ratio of the rate of sites in the two classes. We will plot the signal to noise (observed distance versus δ/s.d.) curve for each set of sites separately, and in combination. Since sites in each set are independent, then $Var[x\delta_a+y\delta_b] = x^2Var[\delta_a] + y^2Var[\delta_b]$ (where *x* and *y* are weights). This model approximates the different rates that might be encountered in a functional region (e.g. for a protein), and so should give some interesting insights to the validity of the practice of excluding third position sites when they appear to approach saturation. The decision to exclude third position sites is sometimes judged to be at the point where d_{obs} of set (a) and d_{obs} of set (b) show a clear curvilinear relationship, e.g. Horai *et al.* 1992. Importantly under this model, we can allocate sites to rate categories a priori. The applicability of the results to real sequences will of course be partially contingent upon how the real sequences evolve (e.g. are we ignoring different substitution processes at third vs first and second position sites).

Figure 4.9 shows the results of these experiments. The results are illustrated with three different ratios for the rate of the fast to slow sites, namely 4, 8 and 16. The average distance was calculated as two thirds of δ_a (the corrected distance for the sites of set *a*) plus one third of δ_b, and the variance of this sum was then estimated as described in the previous paragraph. Clearly the signal-to-noise ratio of the third position sites can "go bad" well before that of the more slowly evolving sites, and in doing so, can ruin the signal-to-noise ratio of the averaged distance (this effect becoming worse as the rates of the two classes of sites become more unequal). Next to each of the plots of signal-to-noise ratio for distances (the right hand plots), is shown a smaller plot of d_{obs} for the slower evolving sites versus d_{obs} of the faster sites (the dark blue lines). In all cases, these plots show a curvilinear relationship. Next, consulting where the average distance has become unreliable (here taken to be when the signal-to-noise ratio of $\delta_{average}$ drops below half that of the first and second positions alone, that is δ_a), it is inferred at what position on the right hand plot this point relates to. This is done by using the red line in the plots to the right, which measure d_{obs} averaged over all sites (the values on the x-axis of the plots on the right) vs. d_{obs} of the slowly evolving sites. Consequently the point where the vertical dotted line (on the plots to the right), intersects the dark blue line indicates when we should seriously consider downweighting or removing the more rapidly evolving set of sites.

(figure next)

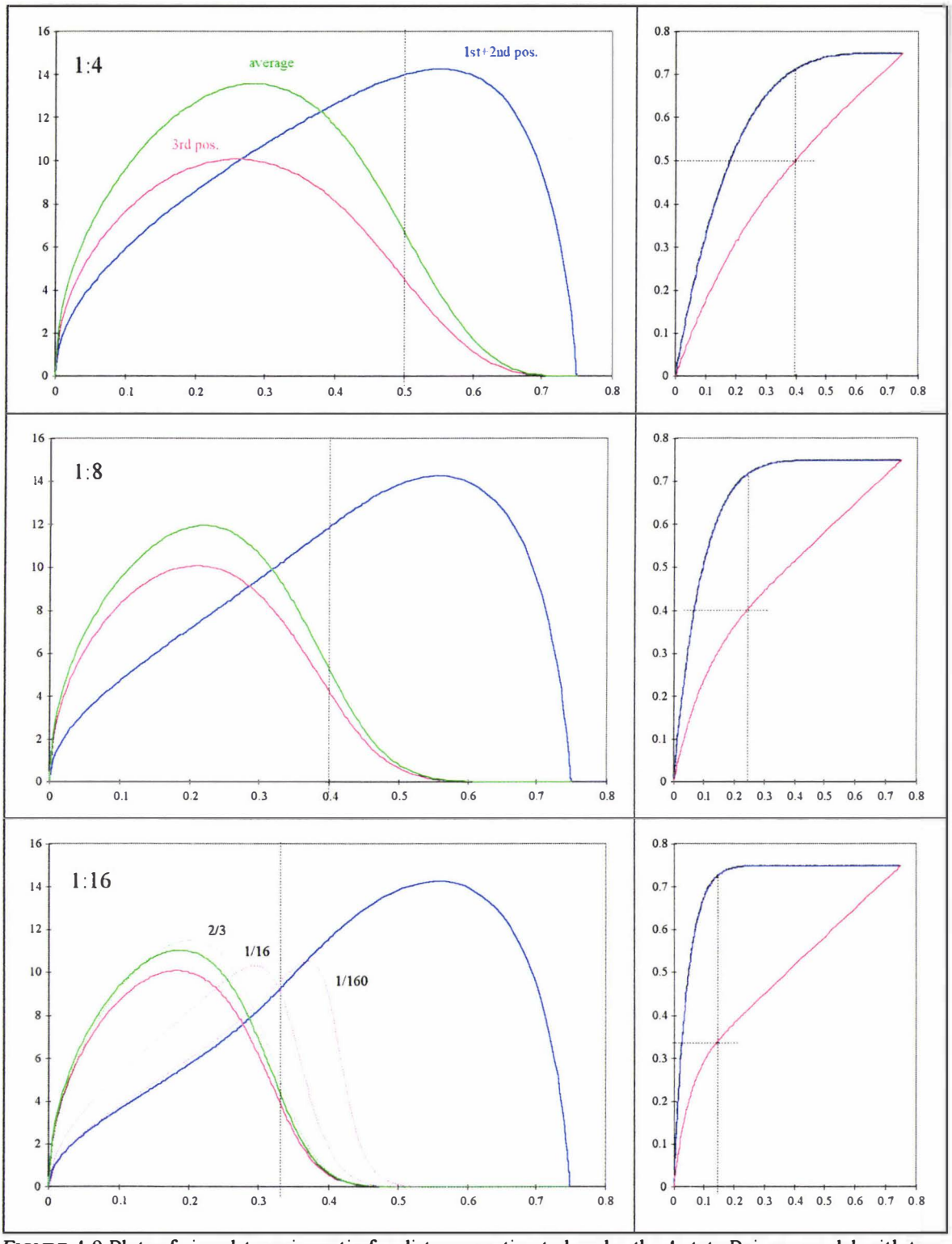


FIGURE 4.9 Plots of signal-to-noise ratio for distances estimated under the 4-state Poisson model with two rate classes (see start of section 4.5.1 for a full description). The ratio of the rate at the third position to that at the first plus second position is 4 in the first row, 8 in the second and 16 in the third. The x-axis of the main (left-hand) plot is observed sequence divergence across all sites. In the smaller plot to the right we plot observed pairwise divergence at first and second position sites (x-axis) vs divergence at third position sites (the blue lines)(as is often done in practice). The red line is a plot of divergence at first and second position sites vs the overall sites sequence divergence. This red line is used to measure the overall sequence divergence value indicated by the dotted line on the main graph, then read across from its value on the y-axis of the second plot, to show the observed sequence divergence at first and second position sites alone. Note, in all cases the point where you would wish to discard the third position sites, coincides

with the middle of the bend in the blue line plots (offering a useful visual guide of when to discard the distance estimate from third position sites). Note that when the maximum signal-to-noise ratio drops to a half its peak value, there have been more than 2.0 substitutions per site on average for sites in that rate class. Finally, the last row shows the effect of downweighting third position sites by a factor of 1/16 and 1/160, then estimating the variance of the average of all sites (the dotted purple lines). Interestingly a subtle down weighting of third position sites by 2/3, yields the dotted green curve which suggests a general improvement over the straight averaging of all sites (the light green curve). Note that while this does allow the averaged distance to have the best signal-to-noise ratio in certain regions, it still does not allow for accurate inference at much higher observed distance values.

One way to try and make the average distance more reliable across a larger range of distances is to downweight the contribution of the rapidly evolving sites. In the last row of figure 4.9, the plot on the left shows the results of such weightings (i.e. $\delta_{weighted} = [2/3\delta_a + y1/3\delta_b]$, where y is the weighing factor of the more rapidly evolving sites). The subtle downweighting with y = 2/3 (the dotted green line) gave good results compared to the standard average (y set to 1) for almost all distances (and with the rate ratios 4, 8, and 16). This weighing factor may be useful when one feels the more rapidly evolving sites do make a positive contribution to estimating the total distance. More severe weightings (1/16 and 1/160) push the optimal signal-to-noise ratio of the rapidly evolving sites to the right (and hence to longer distances). A useful weighting of whole distance matrices to better take into account stochastic error (but not putting emphasis on any particular distances) appears to be,

$$\mathbf{D}_{(a+b+..n)} = m\mathbf{D}_a / \text{s.d.}[\mathbf{D}_a] + m\mathbf{D}_b / \text{s.d.}[\mathbf{D}_b] + ... + m\mathbf{D}_n / \text{s.d.}[\mathbf{D}_n], \qquad (4.5.4\text{-}1)$$

where $\mathbf{D}_x$ is a distance matrix estimated from set of sites x, s.d. is the average standard deviation of distances within that matrix (a matrices scaling factor), while m is a scalar which ensures that the weighting factors sum to one ($m = 1 / \Sigma\, \text{s.d.}(\mathbf{D}_x)$).

Figure 4.9 shows that uniform weighting schemes for whole matrices are clearly a compromise. While they are an improvement over using just the slowly evolving sites for small to moderate distances, they can quickly loose all information at larger distances, and also don't do nearly as well as third position sites alone in estimating shorter distances. Unfortunately, for current tree estimation algorithms based on distances, additivity is required to guarantee consistency. As we prove in appendix 3.5, any sum of single weightings of distance matrices still gives additive distances when these separate distance matrices are combined (given that the original matrices were additive on the same unweighted tree), but different weightings of distances within the same matrix do not make this guarantee. (If so, it would be possible to derive a distance which combined the distance estimates of fast and slow sites using their relative variances at each distance; the resulting signal-to-noise ratio would beat that of any signal-to-noise ratio in figure 4.9). The variable weighting of transversions over transitions (e.g. Schöniger and von Haeseler 1993, appendix 2.6) which can preserve additivity in expectation, is a distinct case where the weighting is occurring between states, and not between sites.

There is, however, a way to take into account the distinct signal-to-noise ratios of all distances, and that is via generalised least squares (GLS) tree selection (this method is described

extensively in chapter 5). To do this, apply GLS tree estimation to the distance matrix for each rate class, and then sum up the difference in GLS scores (sum of squares, SS) for the different trees. This works because GLS takes into account the variance-covariance structure of distances from any single rate class, outputting a measure of fit which is smallest for the best tree. Secondly, this fit measure is additive, so the overall GLS score of T_1 vs T_2 is GLS $SS(T_1(\delta_a) + T_1(\delta_b))$ vs GLS $SS(T_2(\delta_a) + T_2(\delta_b))$, (the best tree being that with the overall minimum GLS SS). Consequently, using this method may offer an advantage by more fully using the information from each separate site class. (Note that ML tree selection on sequences can work in the same way, giving additive likelihoods for the data from different rate classes).

It is interesting that even using statistically efficient estimators like ML or GLS, we may still do better by excluding some of the more rapidly evolving sites from the final analysis. The more rapidly evolving sites can add a lot of stochastic noise to our overall GLS SS statistic, and in the extreme of being completely randomised, they will add just noise and no tree resolving signal. Accordingly, while the difference between trees is expected to become larger as we add more data, some of this data does more to make the difference between trees indistinct than distinct. In the case of ML tree selection from sequences, consider there is a set of sites (x) which clearly demarcate two trees to be tested against each other, and such a test is made (e.g. the bootstrap proportion where tree a is better fitting than tree b, a test approximated by that of Kishino and Hasegawa 1989, both of which are a type of likelihood ratio test). Imagine repeating the same test, but using a set of completely random sites (y, which have the same history but have evolved at a much higher rate). Assuming no systematic errors, then the likelihood difference between the two trees with the second set of sites has an expected value of zero, but potentially a large variance. Adding the likelihood ratio statistics from x and y together will clearly give a likelihood ratio statistic with the same expected mean value, but more variance than the likelihood ratio form set x alone. Clearly, we can potentially increase the resolving power of all tree to tree testing, or tree selection, by being suitably, and objectively selective of the sites being used in an analysis. This can take the form of a variety of weighting schemes.

The findings of the above section have particular relevance to distance and ML methods which integrate rates across all sites (for example the extended Hadamard conjugations of chapter 2). When faced with an analysis, the simplest approach is include all the aligned sites together, then use a correction which approximately accounts for site to site rate heterogeneity (e.g. assume a Γ distribution of rates across sites and optimise the shape parameter k). Unfortunately, for most problems, this will not be the optimal way of resolving a phylogenetic question and being an automatic approach has the danger of becoming a "lazy standard" in contemporary analyses. That is not to say that ML fitting of distributions of rates across sites to observed data does not offer one of the most useful ways of studying this feature of molecular evolution; it clearly does. However this is quite a distinct question to resolving a series of speciation events. The issue of site weighting does not go away, and is something which has yet to be addressed comprehensively in a likelihood frame work. (The program of Olsen 1994 which allows site weighting in ML analyses is a step in this direction. However, it seems unlikely to

have the weighing functions, or the site rate estimations exact. In addition it can easily be fooled by data which comes from a "Felsenstein zone" tree when few taxa are involved, when the integrative methods, while not maximally efficient, are consistent in such a situation). Clearly the method(s) of preference will be partly dependent upon the extent to which the data can unambiguously inform us of a site rate (and not be confounded by the generating tree's edge lengths).

Thus there is no universal answer as to whether to weight or exclude sites. One conclusion however is that if we desire to resolve a specific phylogenetic question then we should evaluate whether it is best to try to combine all the data into one analysis with some weighting, or whether it is better to edit the data to exclude certain sites which will almost certainly add noise. This does pose the immediate challenge, of when we may do better if we can get the "trouble maker" sites out. If we take this direct editing approach, rather than building the weighting into the model, then after such editing, it is of course, still advisable to use a method which integrates over the rates of the remaining sites. Another conclusion of these studies so far, is that we should not expect to be able to reliably estimate relationships for a whole set of very diverse organisms from just one data matrix. Although we generally appeared to do well with the edited rRNA data of Gouy and Li (1989a)(see chapter 3), resolving the archaebacterial question, for example, appeared to require just the most slowly evolving sites.

As mentioned at the beginning of this section, only stochastic error would be considered. With real data the more evolution has occurred, the more the violations of a model's assumptions are likely to lead to significant systematic errors due to the higher degree of "extrapolation" for unobservable changes. Obviously, in a real analysis, data is likely to be legitimately excluded for both reasons. Accordingly, it is important to appreciate that for evaluating very ancient divergences (e.g. the monophyly of the archaebacteria, or the root of the tree of life) discarding at least 50% of all sites as unsuitable for the purpose is not throwing away data, but, rather a necessary step in order to give the analysis meaning. This almost invariably means selecting sites with the slowest rate of substitution and excluding any sites showing a neutral rate of substitution. Unfortunately at present this is not a popular option, nor has much effort been put into estimating which sites are most reliable. For these reasons, the weighting schemes such as those used by Farris (1969) and Penny and Hendy (1985, 1986) are worthy of careful investigation for the purpose of evaluating a sites probable rate of evolution. Studies need to be conducted to compare these, generally computationally fast methods, to those of likelihood (e.g. Olsen 1994) under current models. It can be expected that these "non-model" methods will give both inferior and superior weightings, depending largely upon the likelihood model's assumptions, and the evolutionary processes that generated the data.

The results of this section generalise the question of when to combine data when using inconsistent methods (e.g. Bull *et al.* 1993) to cover consistent methods, and widens the scope of the discussion from one of systematic errors to include stochastic errors.

4.6 NEW 4-STATE HADAMARD CONJUGATIONS TO REDUCE VARIANCE

In this section we show a way to enable the order 4^{t-1} 4-state Hadamard conjugation to make corrections under either the Jukes-Cantor (Poisson model) or the generalized Kimura 2ST model (where the rates of transitions to transversions can be independent on all edges in the tree). So far the 4-state Hadamard conjugation has offered no alternative but to making corrections according to the generalised Kimura 3ST model (see Steel *et al.* 1992, Hendy *et al.* 1994, and extended to a distribution of rates across sites in chapter 2). This simplification of the Hadamard conjugation model has the advantage of reducing the sampling errors in $\hat{\gamma}$. It will also allow us to examine how important the more generalised models are to obtaining a good fit between model and sequence data. Further, it highlights the existence of whole sets of linear invariants under the Kimura 2ST and Jukes-Cantor models, some of which are tree invariants, while others are model invariants (which allow tests of fit of the data to model expectations). These are a different, though related, set of invariants to those which are associated with the order 2^{t-1} Hadamard conjugations of appendix 2.6.

4.6.1 Kimura 2ST and Jukes-Cantor 4^{t-1} Hadamard conjugations

These novel transformations were inspired after using extended order 4^{t-1} 4-state Hadamard conjugations for likelihood calculations (e.g. Waddell and Penny 1995, and chapter 5). Going from tree to sequences, it is easy to set the model to a generalised Kimura 2ST model by making edge weights in $\gamma(T)$ for the two transversions equal (i.e. averaging bipartition transversion entries in the same row of column). Taking this one step further and making the transitions entries in $\gamma(T)$ equal to those of the two transversions gives $\mathbf{s}(T)$ (which are pattern probabilities or likelihoods) under the Jukes-Cantor model. If you examine the $\mathbf{s}(T)$ vectors of the generalised Kimura 3ST model, then you will find that no two are expected to be equal. If two were identical (say $s(T)_i$ and $s(T)_j$) then $E[s(T)_i - s(T)_j] = 0$), this would constitute a linear invariant, and in Steel *et al.* (1993c) we prove that no linear invariants of any kind exist under the generalised Kimura 3ST model (see also Evans and Speed 1993).

Examining the $\mathbf{s}(T)$ vector under the generalised Kimura 2ST model, we observed pairs of equivalent entries (in fact a whole family of linear invariants if one such entry is subtracted from its mate, table 4.6 gives an example). The equivalence of these entries is understandable in terms of a patterns probability, or likelihood. Under the general 12 parameter i.i.d. model the pattern AAGT has a unique likelihood. However, under the more simple Kimura 3ST model, the patterns CCTA, GGAC, and TTCG all have equal likelihood if the root base composition is equifrequency (see section 2.4.1). Constraining the model further, to the generalised Kimura 2ST model, the pattern AAGT is also equal to AAGC (i.e. a transition separates the first two sequences from the third, and a transversion separates the fourth sequence from all others). Thus, what was one pattern under the generalised 12 parameter model has 16 equivalent patterns under the generalised Kimura 2ST model with equifrequency root composition. There is also a carry over of the property from the Kimura 3ST model, that the sum of the four "equivalent" patterns remains unvaried with respect to root base composition i.e. f(AAGT + CCTA + GGAC

+ TTCG) remains the same. So even with unequal root base compositions, there will be pairs of the 4^{t-1} patterns which are equivalent.

There are even more equivalent patterns under the 1-parameter Jukes-Cantor model. For example, the pattern AAGT implies the first two sequences have the same state, the third sequence a second state, and fourth sequence a third state, so AAGT = AAGC = AATG = AACG = AATC = AACT (if the root base composition is in equilibrium). Table 4.6 highlights the examples of six equivalent order 4^{t-1} patterns that occur under this model, with 4 taxa. Thus unique patterns under the Kimura 3ST model become equivalent to other patterns under its submodels, and these equivalent patterns can be thought of as simple permutations of the original pattern. So, without any additional constraints on root base composition or the number of taxa, there will be pairs of elements in the order 4^{t-1} $\mathbf{s}(T)$ equivalent under any Kimura 2ST model, and sets of three and six equivalent entries in $\mathbf{s}(T)$ generated under the Jukes-Cantor model.

If we take t observed sequences and convert them to the 4^{t-1} relative patterns, then apply a Hadamard conjugation, the resultant sequence pattern frequencies have been corrected for multiple changes under the Kimura 3ST model. If we then go through $\hat{\mathbf{s}}$, and make an unweighted average of all the entries which are equivalent under the generalised Kimura 2ST model (to give vector $\hat{\mathbf{s}}_{K2}$), then apply a 4^{t-1} 4-state Hadamard conjugation, the resulting vector, $\hat{\gamma}_{K2}$, is the observed sequence data corrected according to the Kimura 2ST model. If all the equivalent entries under the 1-parameter Jukes-Cantor model are averaged (unweighted), this gives vector $\hat{\mathbf{s}}_{1P}$. Applying the Hadamard conjugation to this vector generates the vector $\hat{\gamma}_{1P}$, which are the sequence pattern frequencies corrected according to the Jukes-Cantor model. All of these averagings and conjugations can be made independent of knowing the true tree.

(table next)

Table 4.6 Pattern probabilities (s) under constraints of the generalised Kimura 3ST model
(Note: To code into binary indexing here use A = 00, G = 10, C = 11, and T = 01, otherwise see section 2.4.1 for the indexing).

A. General patterns, K 3ST							
AAAA	AAAT	AATA	AATT	ATAA	ATAT	ATTA	ATTT
AAAG	AAAC	AATG	AATC	ATAG	ATAC	ATTG	ATTC
AAGA	AAGT	AACA	AACT	ATGA	ATGT	ATCA	ATCT
AAGG	AAGC	AACG	AACC	ATGG	ATGC	ATCG	ATCC
AGAA	AGAT	AGTA	AGTT	ACAA	ACAT	ACTA	ACTT
AGAG	AGAC	AGTG	AGTC	ACAG	ACAC	ACTG	ACTC
AGGA	AGGT	AGCA	AGCT	ACGA	ACGT	ACCA	ACCT
AGGG	AGGC	AGCG	AGCC	ACGG	ACGC	ACCG	ACCC
B. K 2ST non-homogeneous							
0.664863	0.001119	0.005821	0.005791	0.015348	9.98E-05	0.000148	0.008574
0.016557	0.001119	0.000547	0.000415	0.000455	3.84E-05	3.93E-05	0.000251
0.033447	0.000179	0.005821	0.000415	0.000804	8.04E-05	0.000137	0.000448
0.047996	0.000179	0.000547	0.005791	0.002607	1.90E-05	2.84E-05	0.001389
0.061212	0.000115	0.000549	0.001250	0.015348	3.84E-05	0.000137	0.001389
0.004835	0.000115	0.000644	8.94E-05	0.000455	9.98E-05	2.84E-05	0.000448
0.004531	0.000132	0.000549	8.94E-05	0.000804	1.90E-05	0.000148	0.000251
0.072318	0.000132	0.000644	0.001250	0.002607	8.04E-05	3.93E-05	0.008574
C. K 2ST homogeneous							
0.697677	0.001247	0.002876	0.003801	0.005255	3.60E-05	3.31E-05	0.006468
0.017369	0.001247	0.000272	0.000271	0.000156	1.39E-05	1.56E-05	0.000186
0.035058	0.000163	0.002876	0.000271	0.000275	2.87E-05	2.39E-05	0.000336
0.050304	0.000163	0.000272	0.003801	0.000901	6.60E-06	6.41E-06	0.000981
0.063951	0.000122	0.000270	0.000751	0.005255	1.39E-05	2.39E-05	0.000981
0.005061	0.000122	0.000318	5.36E-05	0.000156	3.60E-05	6.41E-06	0.000336
0.004736	0.000142	0.000270	5.36E-05	0.000275	6.60E-06	3.31E-05	0.000186
0.075823	0.000142	0.000318	0.000751	0.000901	2.87E-05	1.56E-05	0.006468
D. 1P model							
0.397306	0.010344	0.021036	0.031444	0.040200	0.003357	0.003124	0.048152
0.010344	0.010344	0.002736	0.002736	0.001446	0.001446	0.001611	0.001611
0.021036	0.002736	0.021036	0.002736	0.002330	0.002728	0.002330	0.002728
0.031444	0.002736	0.002736	0.031444	0.009953	0.000817	0.000817	0.009953
0.040200	0.001446	0.002330	0.009953	0.040200	0.001446	0.002330	0.009953
0.003357	0.001446	0.002728	0.000817	0.001446	0.003357	0.000817	0.002728
0.003124	0.001611	0.002330	0.000817	0.002330	0.000817	0.003124	0.001611
0.048152	0.001611	0.002728	0.009953	0.009953	0.002728	0.001611	0.048152

Notes to table. Part **A**. gives the order 4^{t-1} site patterns in **s**(T). **B**. gives an example of generalised Kimura 2ST model pattern probabilities, without any further constraints. This model has the same rate of type 1 and 2 transversions, but their rate to relative to the transitions is different on each edge. The tree used to generate this data had randomly generated transition frequencies of (0.021435, 0.048754, 0.062356, 0.085327, 0, 0, 0.103645) and transversion frequencies of (0.001567, 0.008323, 0.007825, 0.021047, 0, 0, 0.011444) (where the indexing is that of the second component of the 4-state indexing of Steel *et al.* 1992, Hendy *et al.* 1994, and section 2.4, starting counting at 1). Hence the true tree is T_{12}. Notice the numerous identical entries in **s**(T) under this model. Pairs of transversion bipartitions, and non-bipartition entries have identical values, but transition bipartitions and s_0 are uniquely valued. An example of a linear invariant is the expected value of the difference of the two red entries in column two i.e. $E[\hat{s}_{2,1} - \hat{s}_{3,1}] = 0.000179 - 0.000179 = 0$. **C**. is an example of pattern probabilities under the homogeneous fixed transition to transversion ratio Kimura 2ST model (obtained by setting the tr/tv ratio = 12.453, with edge

weights the same as those given previously for transitions). Notice that no more identical entries have emerged in **s**(T) given this constraint. **D**. shows pattern probabilities under the 1-parameter Poisson model (Jukes-Cantor) model obtained by setting the relative transition and transversion rates equal on each edge (so all changes occur at the same weight as the transitions in the generalised Kimura 3ST example). Notice that many more equivalencies have arisen, there are sets of six identical nonbipartition patterns with identical probabilities (shown in the same colour), sets of three identical bipartition values (underlined), and a single unique s_0 value. Notice that it is these sets of six which break into pairs under the 2ST models, while the sets of three break into a pair and a unique value. Vector **s**(T) was also generated under a stationary Kimura 3ST model (not shown), but no equivalent entries appeared. Putting a Γ distribution of rates across sites ($k = 0.9$) for all these models, exactly the same equivalent entries appeared.

It is also possible to predict how many patterns will be equal in each submodel. Under the generalised Kimura 2ST model, only patterns with a transversion indicated amongst the taxa can give rise to a second pattern, which is equivalent by interchanging the states indicating a transversion. For example, keeping A fixed at the first taxon, if the pattern has either a T or a C in it, switch T's for C's, or C's for T's, or if both C's and T's appear in a pattern they must be switched simultaneously e.g. ACTA, becomes ATCA. Thus no equivalent entries appear in the first column of the Kimura 2ST models of table 4.6 as the only changes indicated there are transitions, but equivalent pairs occur throughout the rest of the **s** vector. Under the Jukes-Cantor model there is only one unique entry, that is s_0. All bipartition patterns can be permuted to three equivalent patterns. Keeping A fixed for the first taxon, a pattern such as ACAC can be permuted to AGAG and ATAT, as all these patterns have the same likelihood on any tree (so we just swap the C to G or T). Once a pattern involves three or more states, then under the Jukes-Cantor model, there will be six patterns with the same likelihood. Taking for example the pattern ACGA, then switching the G to a T now gives two equivalent patterns, the original plus ACTA. Next, switch the original C to G, which means the original G must become a C to a T to be different, so giving another two equivalent patterns, AGCA and AGTA. Lastly, switching the original C to T, so the original G can then switch, gives the equivalent patterns ATGA and ATCA. Note that patterns with four different states shown do not give any more equivalent patterns, since the fourth state has no choice of what it must swap to, as it must remain different to the other three states. Consequently all patterns with more than two states appearing (that is all entries other than the first row, first column, or leading diagonal) permute to $3 \times 2 = 6$ equivalent patterns under the Jukes-Cantor model.

It is important that the same sets of identical values in **s**(T) appear when there is a distribution of rates across sites. Thus these modifications to **s**(T) are also quite valid prior to the use of the extended Hadamard conjugations of chapter 2. A proof of this is sketched at the end of section 4.6.4. We also conjecture, that equivalent entries in **s**(T) will appear under the submodels of Kimura 3ST even if the root distribution is not in equilibrium and each rate class has a different root distribution of nucleotide frequencies. This is easy to prove, since each rate class will independently behave as it does under an i.r. model, so an order 4^{t-1} **s**(T) vector at rate λ_i, is equivalent no matter what the root distribution (see section 2.4.1). Since **s**(T) with unequal rates across sites is just (**s**(T)λ_i), (and in the case of a continuous distribution, this becomes an integral

in the limit), then the overall **s**(T) vector must also remain the same no matter what the root base composition of each rate class.

4.6.2 Linear tree invariants under the Kimura 2ST and Jukes-Cantor models

While the equivalent entries in the **s**(T) vectors under the Kimura 2ST and Jukes-Cantor models suggest many model-invariants, no tree-invariants are immediately obvious. However if you look at the **r**(T) vectors generated under the Kimura 2ST model you immediately see pairs of identical entries, in a very similar pattern to those in the **s**(T) vector. For the Kimura 2ST model, the positions of equivalent entries are the same as in the **s** vector shown in table 4.6 part 2 when viewing the table from side on (that is turn the table so that the last column as you now see it forms the last row). The only difference in terms of where entries are equivalent, is that the first entry $r_{0,0}$ is at the same position as $s_{7,0}$ in the rotated table.

At the **r** level, tree invariants under the Kimura 2ST model become obvious. Under any resolved 4 taxon tree, the last 8 entries in **r** form two sets of four equivalent entries. On tree T_{12}, the equivalent entries are $r_{56} = r_{59} = r_{60} = r_{63}$, and $r_{57} = r_{58} = r_{61} = r_{62}$. Whereas on tree T_{13}, $r_{56} = r_{58} = r_{61} = r_{63}$, and $r_{57} = r_{59} = r_{60} = r_{62}$, while on T_{14}, $r_{56} = r_{57} = r_{62} = r_{63}$, and $r_{58} = r_{59} = r_{60} = r_{61}$. It is easy to see how these equivalent values interrelate if we write them out as,

T_{12}	56	<u>57</u>	<u>58</u>	59	60	<u>61</u>	<u>62</u>	63
T_{13}	56	<u>57</u>	58	<u>59</u>	<u>60</u>	61	<u>62</u>	63
T_{14}	56	57	<u>58</u>	<u>59</u>	<u>60</u>	<u>61</u>	62	63

Elements of **r** in the same type face are expected to be equal for the tree indicated. This then gives rise to a set of tree invariants: For T_{12} the two element tree-invariants are:

$$r_{56} - r_{59} = 0,\ r_{56} - r_{60} = 0,\ r_{59} - r_{63} = 0, \text{ and } r_{60} - r_{63} = 0$$

$$\text{plus } r_{57} - r_{58} = 0,\ r_{57} - r_{61} = 0,\ r_{58} - r_{62} = 0,\ r_{61} - r_{62} = 0.$$

For tree T_{13} the tree-invariants are:

$$r_{56} - r_{58} = 0,\ r_{56} - r_{61} = 0,\ r_{58} - r_{63} = 0, \text{ and } r_{61} - r_{63} = 0,$$

$$\text{plus } r_{57} - r_{59} = 0,\ r_{57} - r_{60} = 0,\ r_{59} - r_{62} = 0,\ r_{60} - r_{62} = 0.$$

While T_{14} has tree-invariants:

$$r_{56} - r_{57} = 0,\ r_{56} - r_{62} = 0,\ r_{57} - r_{63} = 0, \text{ and } r_{62} - r_{63} = 0,$$

$$\text{plus } r_{58} - r_{59} = 0,\ r_{58} - r_{60} = 0,\ r_{59} - r_{61} = 0,\ r_{61} - r_{61} = 0.$$

Of course these invariants are interdependent, as are higher order invariants, e.g. for T_{12} the invariant $r_{56} + r_{59} - r_{60} + r_{63} = 0$, and so on. In order to use these invariants to identify the tree it is probably best to do an appropriate goodness of fit test (taking into account the correlations in **r**, e.g. by generalised least squares, see chapter 5) of how well each set of equivalent values fit the expectation of all being equal. Thus for tree T_{12} we would ask how well $r_{56} = r_{59} = r_{60} = r_{63}$, and $r_{57} = r_{58} = r_{61} = r_{62}$. The tree which relates to the set of entries which are most closely equal is then taken to be the optimal tree. It is also possible to test whether any tree fits this expectation as

well as would be expected if the model was correct. Importantly these same invariants exist when there is a distribution of rates across sites (this was evaluated numerically, and a proof is given at the end of section 4.6.4).

Interestingly, sets of invariants also exist under the constraint of the Jukes-Cantor model, where they are joined by extra tree invariants in **r** (which implies in **s** also since **r** is a linear combination of **s**). With four taxa, the following sets of entries in **r** have identical values under any tree:

{0}, {1, 8, 9}, {2, 16, 18}, {3, 24, 27}, {4, 32, 36}, {5, 40, 45}, {6, 48, 54}, {10, 11, 17, 19, 25, 26}, {12, 13, 33, 37, 41, 44}, {20, 22, 34, 38, 50, 52}, {29, 30, 43, 46, 51, 53}.

However, the following elements of **r** are equal only in the trees indicated,

T_{12}		14	15	21	23							42	47	49	55		57	58			61	62
T_{13}		14	15				28	31	35	39				49	55		57		59	60		62
T_{14}				21	23		28	31	35	39	42	47						58	59	60	61	

and also elements of this complimentary set will only be equal under the indicated tree,

T_{12}	7					28	31	35	39					56			59	60				63
T_{13}	7			21	23						42	47			56		58			61		63
T_{14}	7	14	15										49	55	56	57					62	63

Notice how the these sets are different for each tree, and it is from these elements that invariants are generated. For example, an invariant of T_{12} alone (of the binary trees) is $r_{14} - r_{21} = 0$. As with the Kimura 2ST model, these invariants still exist when there is a distribution of rates across sites. There are so many interrelated invariants here that the most practical (and statistically integrated) approach to picking a tree, would again appear to be to evaluate which set of tree dependent **r** values come closest to being equal (when doing this test it would be appropriate to drop elements 7, 56, and 63 as they do not discriminate between trees).

As yet we have not explored how the invariants under the Kimura 2ST model relate to Lake's (1987) 'evolutionary parsimony' invariants, but we suspect that with four taxa they will be equal. An important discovery here is that similar sets of invariants exist with more than four taxa (this can be seen looking at **r** generated after the Hadamard transform), and these may be an equivalent to an extension of 'evolutionary parsimony' to many taxa. Further study of this question is required. Presently, we do not see these invariants as being the most reliable way to pick a tree, but their existence is interesting to note and they warrant further study (which is beyond the scope of this thesis). Our skepticism for towards their statistical efficiency in tree selection may be summarised as follows:

(1). They are a strict subset of entries in **r**, and by using these alone we suspect information important to tree selection is being discarded.

(2). These invariants are probably much inferior to ML for tree selection. Assuming these invariants are like those of Lake (1987) and discard information in a similar way, then it has been shown by Navidi *et al.* (1991) using theoretical calculations, plus Jin and Nei (1990) and Hillis *et al.* (1994) in simulations, that these invariants are generally much less statistically efficient than likelihood, parsimony, or distance based methods. Some of these studies were repeated by myself and Dr David Swofford, modifying an earlier program to generate data with an unequal distribution of rates across sites. We were concerned that earlier studies had perhaps been unfair towards 'evolutionary parsimony' since its strength is that the distribution of rates across sites does not need to be specified for it to be consistent. Despite much searching of the Kimura 2ST four taxon parameter space, unless an alternative popular method was going to be inconsistent in estimating the unweighted tree (or very close to the boundary of this situation), it did (usually strikingly) better than 'evolutionary parsimony' in all cases. Such studies now need to be made with tree selection using both the Kimura 2ST and Jukes-Cantor model linear invariants for any number of taxa.

We now return to the theme of estimating the covariance matrix of sequence data transformed under Kimura 3ST model submodels.

4.6.3 Calculating the covariance matrix of $\hat{\gamma}_{K2}$ and $\hat{\gamma}_{1P}$

The averaging of entries in $\hat{s}$ reduces the sampling errors of these entries, and their corresponding entries in $\hat{\gamma}$. If we average two model equivalent entries in $\hat{s}$, then under the i.i.d. assumption, the variance of the average is equal to approximately 1/2 the variance of the unaveraged entries. Consequently the standard deviation has been reduced to approximately $1/\sqrt{2}$ or ≈ 0.7 of its previous value). If there are n entries averaged, the variance of these entries in $\hat{s}$ is reduced to approximately 1/n (and hence their s.d. is reduced to $1/\sqrt{n}$). We say reduced by approximately 1/n because the exact variance is, of course, a scaled binomial. The exact reduction will vary with the expected size of entries in $\hat{s}$. For example, with a sequence of c sites the variance of entry s_i with expected size 0.05 is $s(T)_i(1-s(T)_i)/c = 0.05\times(1-0.05)/c = 0.0475/c$. If $s(T)_i$ is averaged with n other equivalent patterns, its variance becomes $(1/n)ns(T)_i(1-ns(T)_i)/c$, which for $s(T)_i = 0.05$ and $n = 2$, gives its variance as 0.0225/c (which is smaller by $[1-\{1-(1-ns(T)_i)/(1-s(T)_i)\}]\%$). Similarly, the covariance of the pattern s_i for set k of equivalent entries, with another pattern s_j from set l of equivalent entries is $\left(\sum_{i\in k} s_i\right)\left(\sum_{j\in l} s_j\right)/\left(n_k n_l c\right)$. This number is larger than the covariance without averaging (if s_i and s_j are from the same set of equivalent entries, then their covariance increases to the size of their variance). So while we see a more than two fold reduction in variance, this is partly counter balanced by an increase in the negative covariances of entries in $\hat{s}$. Another important property for hypothesis testing, is that averaged entries in $\hat{s}$ will result in entries in $\hat{\gamma}$ with marginal distributions which more closely approximate the normal distribution. Many of the entries in $\hat{\gamma}$ with the largest factor of variance reduction (up to six under the Jukes-Cantor model) will be model invariants, which should make it easier to detect model violations.

To calculate the covariance matrix of $\hat{\gamma}_{K2}$ ($\hat{\gamma}$ under the Kimura 2ST model), it is easiest to start with a condensed form of $\hat{\mathbf{s}}_{K2}$. As discussed in the previous paragraph, it is easy to get the variance of averaged entries in $\hat{\mathbf{s}}_{K2}$, but bookkeeping must be kept in order to recalculate all the new covariances after the averaging takes place. Here we describe a simple way to ensure this. Start by generating a vector $\hat{\mathbf{s}}_{K2C}$ (the subscript $_C$ here standing for condensed form) which has the proportions of sets of sites showing each of the unique patterns that exist under the K 2ST model (thus for 4 taxa $\hat{\mathbf{s}}_{K2C}$ will have 1 + 7 + 56 / 2 = 36 entries). Under the i.i.d. assumption entries in this vector are distributed as a scaled multinomial distribution, with variances and covariances given by equations 4.2.2-1 and 4.2.2-1 respectively. Accordingly, the covariance matrix of this condensed vector, $\mathbf{V}[\hat{\mathbf{s}}_{K2C}]$, is calculated.

Unfortunately it is not possible to combine rows of $\mathbf{H}$ and still effect a fast Hadamard transform of the condensed vector $\hat{\mathbf{s}}_{K2C}$ (Tolimieri *et al.* 1989), so we expand $\hat{\mathbf{s}}_{K2C}$ back out to 4^{t-1} entries. As $\hat{\mathbf{s}}_{K2C}$ is expanded into $\hat{\mathbf{s}}_{K2}$, divide each entry in $\hat{\mathbf{s}}_{K2C}$ up by the number of patterns (n) that it generates in $\hat{\mathbf{s}}_{K2}$ and apportion these to their places in the indexed 4^{t-1} vector. The variance of each entry in $\hat{\mathbf{s}}_{K2}$ thus becomes $\mathbf{V}[(\hat{\mathbf{s}}_{K2C})_i]/n$. The covariance of entry s_i and s_j in vector $\hat{\mathbf{s}}_{K2}$ is equal to $\text{cov}(s_k, s_l)/(n_k \times n_l)$, $k \neq l$, in vector $\hat{\mathbf{s}}_{K2C}$, where n_k is the number of patterns that entry k of $\hat{\mathbf{s}}_{K2C}$ generates in vector $\hat{\mathbf{s}}_{K2}$ (and likewise for l). Of course if pattern k and l are equivalent in $\hat{\mathbf{s}}_{K2C}$, then their covariance in $\hat{\mathbf{s}}_{K2}$ is equal to their variance (e.g. see Stuart and Ord 1987, section 10.6). Completing these calculations gives the covariance matrix $\mathbf{V}[\hat{\mathbf{s}}_{K2C}]$. It turns out that the covariances obtained by dividing $\text{cov}(s_k, s_l)$ by $n_k \times n_l$, have the same expected value as $\text{cov}(s_i, s_j)$ when s_i and s_j are a pair of entries in the order 4^{t-1} vector of a Kimura 2ST model, but the sampling variance of this estimate covariance is reduced by averaging. The variance of the estimated covariances of the averaged entries decreases by approximately order $(n_l \times n_k)$ from their value in the 4^{t-1} vector without averaging. This is important when making statistical inferences as we have shown by simulations (section 4.4.2) that estimated covariances and especially correlations are themselves highly variable. Of course some entries in $\hat{\mathbf{s}}_{K2}$ will now have a correlation of one, if $\text{var}(s_i) = \text{var}(s_j)$. This complete correlation of entries in $\hat{\mathbf{s}}_{K2}$ will then find its way through the transformation to be expressed as the complete correlation of the same entries in $\hat{\gamma}$.

By the same arguments used in sections 4.2 then,

$$\mathbf{V}[\hat{\gamma}_{K2}] = \mathbf{H}^{-1}(d(\mathbf{H}\mathbf{V}[\hat{\mathbf{s}}]\mathbf{H}))\mathbf{H}^{-1} \qquad (4.6.3\text{-}1)$$

where $\mathbf{H}$ is a Hadamard matrix of $4^{2t-2} = 8^{t-1}$ entries, and d denotes the delta method approximation of variances, which for the i.r. Kimura 2ST model is given by equations (4.2.4-1 and 4.2.4-2). By altering the delta function (d) in equation 4.6-1, it is straightforward to estimate $\mathbf{V}[\hat{\gamma}_{K2}]$ if corrections are made according to some specific distribution of rates across sites (see section 4.5.1). So to recap, it is important to keep a track of all the changing covariances once the averaging of entries in the order 4^{t-1} vector begins, and the easiest way to do this is to start with the covariance matrix of the unique patterns only and then expand this out to be the covariance matrix of a vector of 4^{t-1} entries necessary to implement the Fast Hadamard

transform. This same type of procedure is also used to estimate the covariance matrix ($\mathbf{V}[\hat{\gamma}_{1P}]$.) of the 4^{t-1} $\hat{\mathbf{s}}$ vector after averaging equivalent entries in the Jukes-Cantor model, to give the condensed vector $\hat{\mathbf{s}}_{K2C}$ (which has just 1 + 21/3 + 42/6 = 15 entries).

Note, that while it is simple to average entries that would be equivalent under the Kimura 2ST or the Jukes-Cantor model after applying the Hadamard conjugation, this does not force the corrections to the more restricted models. It would, however, reduce these entries sampling variance which could be useful for tree selection. If there were no non-linear transformation, averaging in γ would be the same as averaging in **s**. Generally we expect that averaging before the non-linear transformation will do more by keeping the increased variance at this step to a minimum. It remains to be evaluated how much improvement forcing the corrections to the Kimura 2ST or Jukes-Cantor models will offer in typical applications, where the advantages of reduced sampling error would be contrasted with the greater generality of the Kimura 3ST transformations.

4.6.4 Testing difference in fit between either $\hat{\mathbf{s}}_{K2}$, $\hat{\mathbf{s}}_{1P}$, and $\hat{\mathbf{s}}_{K3}$

Here we describe a test for a significant difference in fit between $\hat{\mathbf{s}}_{K2}$ and $\hat{\mathbf{s}}_{K3}$ (and by like argument between either $\hat{\mathbf{s}}_{1P}$ and $\hat{\mathbf{s}}_{K3}$ or $\hat{\mathbf{s}}_{K2C}$ and $\hat{\mathbf{s}}_{1P}$). It uses the whole family of linear invariants which exist under the generalised Kimura 2ST model. The first step is to convert these **s** vectors back into **f** vectors of observed counts, by multiplying by c, the sequence length. Next we define the test statistic,

$$X^2 = \sum_i \frac{(f_{(K3)_i} - f_{(K2)_i})^2}{f_{(K2)_i}}, \qquad (4.6.4\text{-}1)$$

where i are just those entries which have different expected values under the generalised Kimura 3ST and 2ST models (that is excluding the transition bipartitions, but including s_0, which is the sum of the remaining entries). Asymptotically (as $c \to \infty$) then under the null model (the generalised Kimura 2ST in this case), X^2 converges to a χ^2 distribution. The degrees of freedom of this distribution will be x -1 - {no. of parameters estimated from the data}, and where x will be the number of distinct entries between the two models, which in the case of the Kimura 3ST versus the Kimura 2ST model = $(4^{t-1}-2^{t-1})$. The parameters being estimated are the expected values of equivalent cells under the more restrictive model (so for the Kimura 2ST model, there are $(4^{t-1}-2^{t-1})/2$ pairs of cells, leaving an asymptotic $(4^{t-1}-2^{t-1})/2$ - 1 degrees of freedom). As already mentioned in chapter 1, the vector **f** often suffers from extreme sparseness (very low expected frequencies in most cells). If the usual guides to invoking a chi-square approximation for the distribution of the X^2 statistic do not hold (e.g. more than 75% of cells have counts of four or more, with no cells having counts of less than 1), it is strongly advised to either: (a) use a simulation to estimate the sampling distribution of X^2. (b) group cells (as long as they do not have the same expected values under the Kimura 2ST model)(see Stuart and Ord 1990, p. 1172 for advice on grouping cells). Of course, rejection of the null hypothesis that the data comes from a simpler model does not imply, for example, that $\hat{\gamma}_{K3}$ will be a more useful starting point

than $\hat{\gamma}_{K2}$ for either tree selection or biological inference. Nor does rejection of the null hypothesis imply that the Kimura 3ST model fits the data adequately.

Lastly, under the assumption of equal base frequencies at the root of the tree, we can test the fit of data to Kimura 3ST model expectations. If this model holds, we expect the frequency of sets of 4 patterns to be equal. Specifically patterns such as AAGT, CCTA, GGAC and TTCG should all have equal frequency. One implication of this finding is that if we multiply the ordered 4^t vector of observed frequency patterns for four taxa by **H**, then the expectation is that 192 entries will have value zero, while 64 entries may take positive values. Another way of testing this expectation is that the sum of all patterns with an A in the first position, should be equal (within sampling error) to the sum of all patterns with a C in the first position, should be the same as the sum of all patterns with G in the third position, should be the same as the sum of all patterns with a T in the first position. In actuality, this test will often reject the null hypothesis that the data come from this model. The generalised Kimura 3ST model also allows for unequal base composition at the root. To test this generalised model it is sufficient to test the fit of the predicted 4^{t-1} patterns, against their observed frequencies (as we do in chapters 5 and 6).

However rejection of the fit of data to model (especially by a powerful test), does not necessarily mean that a model is not useful. Such a model may incur some systematic errors, but it may also be less sensitive to sampling errors, making it a more useful model in practice. Hence while fit of data to model may be helpful in inferring some characteristics of sequence evolution, by itself it does not tell us which model will be most reliable for phylogenetic inferences.

Capitalizing on the potential usefulness of the results presented in this section of the thesis, Hendy and Penny (1996) have described an indexing, and counting of the number of linear invariants under these submodels of the Kimura 3ST models. Their results should be helpful to the programming of these methods. They have also counted and indexed the set of entries in **r** which are equivalent under the molecular clock hypothesis for the Kimura 3ST and submodels. Averaging of entries in both **s** and **r** may then allow us to constrain Hadamard conjugations to obey not only submodels of Kimura 3ST but also expectations of a molecular clock. Calculating the covariance matrix of γ in such cases would use the results of these sections to estimate $\mathbf{V}[\hat{\mathbf{s}}]$, then results in the distance Hadamard sections (later) to reestimate $\mathbf{V}[\hat{\mathbf{r}}]$ after averaging the clock equivalent entries in **r** (followed by the remaining steps described in section 4.2 to obtain $\mathbf{V}[\hat{\gamma}]$). This should then allow for tree selection under the constraint of a molecular clock by using standard tree section criteria (see chapter 5) to select a tree from the clock constrained γ vector.

An important new result in Hendy and Penny (1996), is to prove that all linear invariants at the level of **s**(T) under Kimura 3ST submodels must also show up as equivalent entries at the level of **r**(T). This allows us to describe a simple proof that the same equivalent entries in **s**(T) will show up under the various submodels of the Kimura 3ST model, with or without unequal rates across sites. Firstly, note that unequal rates across sites have no impact upon the transformation from γ(T) to ρ(T). It is also true that the only linear relationships (including those

which generate the equivalent entries in **s**(T) under the i.r. submodels) which can generally cross the non-linear transform between **r** and ρ, are of the form $x = y$. Thus, as long as any monotonic function is used going from ρ(T) to **r**(T)(which includes any distribution of rates across sites), the same equivalent entries must appear in **s**(T) with or without a distribution of rates across sites. That the only linear relationships which can generally pass from the additive ρ(T) to the non-additive **r**(T) are of the form $\rho_i = \rho_j$, so $r_i = r_j$, also explains why linear invariants present in ρ(T) under a homogeneous Kimura 2 or 3ST model, do not show up in **r** or **s** level. Thus the only ways at present to take advantage of this reduction in model parameters (and also sampling variance) is via the order 2^{t-1} 4-state Hadamard conjugations of appendix 2.6, or via maximum likelihood and related types of tree selection (see chapter 5).

Some similar results on tree invariants under the Jukes-Cantor model are reported in Fu and Steel (1995), dealing especially with counting their number and indexing.

4.6.5 Reduced variance and bias by using new pathlength transformations

Recently Tajima (1993a) suggested a new class of distance transformations, which have reduced bias relative to the standard Jukes-Cantor and Kimura 2ST distance transformations (we call them reduced bias or *rb* estimators as there remains an often tiny, but theoretically expected, amount of bias). They are based on replacing the logarithmic function with a Taylor series type approximation, taken to c terms. Tajima (1993a) derived these estimators for 4-state pairwise distance transformation. In appendix A4.2 we show they can also be applied to 2 and 4-state pathlength transformation, so giving rise to a new variant of the Hadamard conjugation when going from sampled data, $\hat{\mathbf{s}}$, to $\hat{\boldsymbol{\gamma}}$,

$$\hat{\boldsymbol{\gamma}} = \mathbf{H}^{-1}(rb(\mathbf{H}\hat{\mathbf{s}})), \qquad (4.6.4\text{-}1)$$

where *rb* refers to equation A4.2.1-1, of appendix 4.2. Initially these estimators were seen as a potentially useful answer to the question of what to do when the ML estimate of a pathset length is infinity (i.e. the argument to the logarithmic transform is inapplicable, which occurs when the $r_i \leq 0$). These *rb* estimators will transform any possible r_i value, from 1 to -1, with no need of special treatment.

Further study of the properties of these new pathlength estimators revealed they have the more important property of reduced sampling error, relative to the standard logarithmic estimator (which as previously mentioned is also a maximum likelihood estimate under the model). This finding has special importance, it is the first time in my knowledge that it has been shown, that a model based genetic distance transformation to restore additivity can have substantially reduced sampling variance relative to a maximum likelihood (ML) estimator. The reduction in the root mean square error (RMSE) can be up to 15%, which is roughly equivalent to increasing the sequence length by $(1 - (1/0.85)^2) \times 100\% \approx 38\%$ (see appendix 4.2). Interestingly *rb* estimators tend to make the most improvement on what would often correspond to longer pathset lengths in a tree. A study of the properties of these estimators is given in appendix 4.2. (These results are kept separate to the main text, so as not to detract from the main theme, but are major results).

This new finding of reduced variance properties, suggests that the derivation of similar estimators to infer pairwise distances under more general models (and with a distribution of rates across sites) could be of great use for phylogenetic inference when working with either pairwise distances or Hadamard conjugations.

4.7 SAMPLING ERRORS OF $\hat{\gamma}_D$, THE DISTANCE HADAMARD

In this section we look at the statistical structure of an interesting variant of the Hadamard conjugation known as the "distance Hadamard transform" or just the "distance Hadamard" (Hendy and Penny 1993). As its name suggests it uses only distances in order to construct a vector of partition weights, and can therefore use more types of both data and transformation than the full conjugation. Prior to this work nothing was known of its statistical properties, except that under the model (with additive distances) it is consistent (in that it would return the γ vector of the tree used to generate the distances). Since it deals only with distances (transformed or not) it does not become more complicated as our data gains more states (c.f. the Hadamard conjugation), and so essentially remains an analogue of the 2-state Hadamard conjugation. Results in chapter 5 show that $\hat{\gamma}_D$ has some desirable properties with regard to tree selection, which were predicted from the statistical properties described below.

Despite there being no known tree-independent formulae for pathset lengths under models more general than Cavender's 2-state or the generalised Kimura 3ST with 4-state data, we can still infer a type of gamma vector ($\hat{\gamma}_D$) for more general models. We do this by inferring the pathset lengths of quartets in ρ from pairwise distances using the four point metric (Buneman 1971). For pathsets with 6 or more tips, a natural extension of the four point metric is used to infer their size (Hendy and Penny 1993, and described below). The vector $\hat{\gamma}_D$ will be consistent (in recovering a vector description of the true tree), if the distance corrections are themselves consistent (that is, they have converged to true tree additive distances as $c \rightarrow \infty$). Dissimilarity measures such as DNA hybridisation and immunological distances can be used in this version of the Hadamard conjugation. For pathsets of order greater than 1 (i.e. 2 or more end points or pairwise distances),

$$D(E_k) = \min_{i \neq j \in E_k} (d_{ij} + D(E_k - \{i,j\})), \qquad (4.7\text{-}1)$$

where $D(E_k)$ is the sum of distances with unique end points within the even order set of taxa, k, (Hendy and Penny 1993). We may make this calculation by fixing any i then checking each remaining value for $j \in k$. Then, E_k is replaced with $E_{(k-ij)}$ and this loop is repeated until there are just two elements left in set k. Using this method to infer all pathset lengths from just pairwise distances the vector $-\rho_D/2$ is constructed. The vector $\hat{\gamma}_D$ is then calculated as $-2\mathbf{H}^{-1}\text{-}\rho_D/2$. It is interesting to note that given just pairwise distances and the knowledge that this is a pathset

length on a tree, and assuming the model to be correct, then the minimum of a pair of distances appears to be the maximum likelihood estimate of that pathset length on the true tree. This is because the distances are expected to have zero correlation (see section 4.2.4), and the expected sum of distances on a path is expected to be minimal. A similar argument holds for pathsets with more than four tips.

4.7.1 The covariance matrix of $\hat{\rho}_D$

We begin by calculating the entries in $\mathbf{V}[\hat{\rho}_D]$ relating to their pairwise distances, then from these calculate the variances and covariances of the inferred pathset lengths (which are just unweighted sums of distances). The variances of transformed distances may be obtained by the delta method as discussed earlier. If no transformation is made (e.g. the norm with DNA hybridisation distances) then we use the estimated variances and covariances of the observed data. This gives us all the entries in a covariance matrix of the pairwise distances, which is a subset of all entries in $\mathbf{V}[-\hat{\rho}_D/2]$. Next we calculate the variance of each of the higher order pathsets. We may do this in two different ways. Since the inferred value of a quartet (pathset order two) is just the sum of two distances, then,

$$\mathrm{Var}(D_{ijkl}) = \mathrm{Var}(d_{ij}) + \mathrm{Var}(d_{kl}) - 2\mathrm{Cov}(d_{ij},d_{kl}), \qquad (4.7.1\text{-}1)$$

where D_{ijkl} is the minimal sum of distances with the listed end points, and "Var" and "Cov" refer respectively to the variances and covariances of the distances making up the sum. For example if D_{ijkl} is estimated as the sum of $d_{ij} + d_{kl}$ (with variances 0.2 and 0.1, respectively, and covariance 0.004), then $\mathrm{Var}(D_{ijkl}) = 0.2 + 0.1 - 2 \times 0.004$. The variance of higher order pathsets, for example a pathset with six tips estimated as the sum of 3 distances becomes,

$$\begin{aligned}\mathrm{Var}(D_{ijklmn}) = {} & \mathrm{Var}(d_{ij}) + \mathrm{Var}(d_{kl}) + \mathrm{Var}(d_{mn}) \qquad (4.7.1\text{-}2)\\ & - 2\mathrm{Cov}(d_{ij},d_{mn}) - 2\mathrm{Cov}(d_{kl},d_{mn}) - 2\mathrm{Cov}(d_{ij},d_{mn}).\end{aligned}$$

For still larger pathsets we sum the variances of the component distances, and subtract $-2\mathrm{Cov}(d_{ij},d_{kl})$ for all possible pairs of $d_{ij} \neq d_{kl}$ from the set of labels k that define that pathset.

The second method of estimating $\mathrm{Var}(D_{ijkl})$ uses the expectation that if d_{ij} and d_{kl} are indeed non-intersecting paths, then their covariance will be zero under an i.i.d. model (since they are disjoint sets). Thus we would estimate $\mathrm{Var}(D_{ijkl}) = 0.2 + 0.1$ (we treat the last term as error due to finite sample size). Later we use a simulation to check that there is no obvious bias introduced into $\mathbf{V}[\hat{\gamma}_D]$ as a result of this assumption. Under an i.i.d. tree model, then any non-zero value of $\mathrm{Cov}(d_{ij},d_{kl})$ (where $d_{ij} + d_{kl}$ is the minimum sum of pairwise distances with tips ijkl) due to either sampling error or a violation of the models assumptions (the crucial assumptions being that there has been i.i.d. evolution on a tree, and that the pairwise distances are additive upon the tree). With other types of distances data, for example pairwise DNA hybridisation distances, then non-zero covariances can legitimately be included as they are probably the results of experimental errors (which are not necessarily independent).

We now look at estimating the covariances amongst pathsets in $-\hat{\rho}_D/2$. By the first method the covariance of D_{ijkl} with d_{mn} is,

$$\text{Cov}(D_{ijkl}, d_{mn}) = \text{Cov}(d_{ij},d_{mn}) + \text{Cov}(d_{kl},d_{mn}) + \text{Cov}(d_{ij},d_{kl}), \qquad (4.7.1\text{-}3)$$

(given that d_{mn} is not a member of the set of distances making up D_{ijkl}). For pathsets of order > 2 (that is 6 or more end points), then we sum the covariance of d_{mn} with all the distances making up D_{ijkl}, and then subtract the pairwise covariances within the set of distances comprising D_{ijkl}. If d_{mn} is a member of the set of distances in D_{ijkl} (for example $d_{mn} = d_{ij}$) then the covariance of d_{ij} with itself is Var (d_{ij}), so

$$\begin{aligned}\text{Cov}(D_{ijkl}, d_{ij}) &= \text{Var}(d_{ij}) + \text{Cov}(d_{kl},d_{ij}) + \text{Cov}(d_{ij},d_{kl}), \qquad (4.7.1\text{-}4)\\ &= \text{Var}(d_{ij})\end{aligned}$$

We may also use the second method (that non-intersecting paths have covariances of expected value zero under an i.i.d. model) when inferring covariances. For any d_{mn} (that is irrespective of its being a part of the sum of distances making up D_{ijkl} or not) we have the theoretical expectation that all covariances within the set of distances comprising D_{ijkl} will be zero. Therefore Cov(X, Y) = Σ Var(d_{xy}) (where X and Y are estimated according to equation 4.7-1, and d_{xy} is any distance which is a component of both X and Y). Using this formula speeds up the calculation of $V[-\hat{\rho}_D/2]$ considerably when working with larger sets of sequences.

The last step in calculating $V[\hat{\gamma}_D]$ is equivalent to that used in section 4.2.5 to estimate $\mathbf{V}[\hat{\gamma}]$. Since $\hat{\gamma}_D = \mathbf{H}^{-1}\text{-}2(-\hat{\rho}_D/2) = -2(\mathbf{H}^{-1}(-\hat{\rho}_D/2))$ and given that $\mathbf{H}^{-1}$ is an orthogonal and unweighted linear transform, then

$$\mathbf{V}[\hat{\gamma}_D] = \mathbf{H}^{-1}\mathbf{V}[\hat{\rho}_D]\mathbf{H}^{-1} = 4(\mathbf{H}^{-1}\mathbf{V}[-\hat{\rho}_D/2]\mathbf{H}^{-1}) \qquad (4.7.1\text{-}5).$$

A worked example of calculating $\mathbf{V}[\hat{\rho}_D]$ is given in table 4.7, based on the same model data used in tables 4.1 to 4.5. The major finding is that the variance of the longest pathset estimate, the 4-tipped pathset ($\hat{\gamma}_{D7}$) has decreased by nearly 2/3rds. This finding (and theory) suggests that the reduction of variances in $\mathbf{V}[\hat{\rho}_D]$ relative to $\mathbf{V}[\hat{\rho}]$, should be especially useful when the longest pathsets are larger than the largest pairwise distance. In addition, the variance of the pathset length will be minimised when each of the distances summed up to estimate it are of identical size. These two conditions will generally be best met in large diameter trees, without large internal edges; just the type of tree that is generally most difficult to recover. In addition, higher order pathset lengths will tend to be longest on trees with an evenly branched topology (tree shape), rather than a caterpillar topology (a long back), assuming edge lengths are random variables and the sum of edge weights in each type of tree is equal.

It is also interesting to note that in this example, the **r** entries for the pair of distances estimating the pathset with four tips was 0.63, while that of the direct inference of the 4-tipped pathset was 0.4. Consulting figure 4.6b we see that both estimates have a similar signal-to-noise ratio, but the signal-to-noise ratio of the sum of the two distances improves by approximately one over square root of 2/3 or 1.22 (22 %). This is a considerable improvement. Since 0.4 is near the optimal signal-to-noise ratio for a single pathset length estimate, then generally as the 4-tipped pathset length increases, the sum of distances will be an even more accurate estimate of its true value (model assumptions being meet), while if it is larger (i.e. the pathset length is shorter) then the sum of pairwise distance won't offer such an advantage.

Table 4.7 The covariance matrix $V'[\hat{\rho}_D]$ for the data used to estimate $V'[\hat{\rho}]$ earlier in table 4.3

Index		0	1	2	3	4	5	6	7
Pathset		{0}	{1,4}	{2,4}	{1,2}	{3,4}	{1,3}	{2,3}	{1,2; 3,4}
ρ(T)		.00	.50	.15	.45	.45	.85	.50	.90
$V'[\hat{\rho}]$ Index									
0	0	.000	.000	.000	.000	.000	.000	.000	**.000**
1	d_{14}	.000	1.718	.221	1.226	.105	1.460	.105	**1.331**
2	d_{24}	.000	.221	.350	.105	.105	.105	.221	**.210**
3	d_{12}	.000	1.226	.105	1.460	.000	1.226	.105	**1.460**
4	d_{34}	.000	.105	.105	.000	1.460	1.226	1.226	**1.460**
5	d_{13}	.000	1.460	.105	1.226	1.226	4.474	1.460	**2.452**
6	d_{23}	.000	.105	.221	.105	1.226	1.460	1.718	**1.331**
7	d_{12+d34}	*.000*	*1.460*	*.221*	*1.460*	*1.460*	*3.953*	*1.460*	*5.050* \ **2.950**

All variance and covariance values not in bold are the values one would see in the Hadamard conjugation from 2-state sequences (i.e. identical to those in table 4.3). The values of the last row, marked in italics, are the variance (diagonal element) and covariances of the pathset with 4 tips when it is estimated directly, as in the Hadamard conjugation. The values in bold are the equivalent entries when the 4-tipped pathset is estimated from pairwise distances. Underlined are the two values in each row which are summed to estimate the covariances and variance of the 4-tipped pathset estimated with pairwise distances. Notice that while some of the covariances have decreased slightly, the major change is in the variance of the 4-tipped pathset, which is now 2/3rd's as large as in the Hadamard conjugation.

We now look in more detail at the last point of the previous paragraph, and evaluate the signal-to-noise ratio of many-tipped pathsets estimated as a sum of pairwise distances. For illustrative purposes we have chosen to evaluate a 4-tipped pathset. When using pairwise distances we have assumed that the two distances making up the pathset length are either identical (the optimal situation) or else occur in a ratio of 4:1 (less optimal). It is hard to know what the relative size of the two "halves" of a 4-tipped pathset are expected to be when considering real data, but we expect it would typically fall between these two values. Figure 4.10 shows the results of our study. Clearly, the sum of distances never does worse than the direct inference method, although the difference is minor for pathsets of up to 0.2 substitutions per site. However beyond this value the sum of distances performs better and better, reaching its peak signal-to-noise ratio at somewhat larger pathset lengths than the direct inference equation. The point at which the sum of distances (both distances equal) achieves a signal-to-noise ratio twice that of the direct inference method is marked, and this occurs at δ slightly less than 1. In the case of pathsets with more tips, then the signal-to-noise ratio of a sum of distances tends to do even better than the direct inference method. This suggests that when simultaneously analysing both more sequences and more divergent sequences, then the ratio of the statistical efficiency of $\hat{\gamma}_D$ to $\hat{\gamma}$ (with respect to estimating $\hat{\gamma}(T)$) will get larger. This of course assumes that the model's assumptions hold, and that the correlation structure of $V[\hat{\rho}_D]$ does not preclude this result.

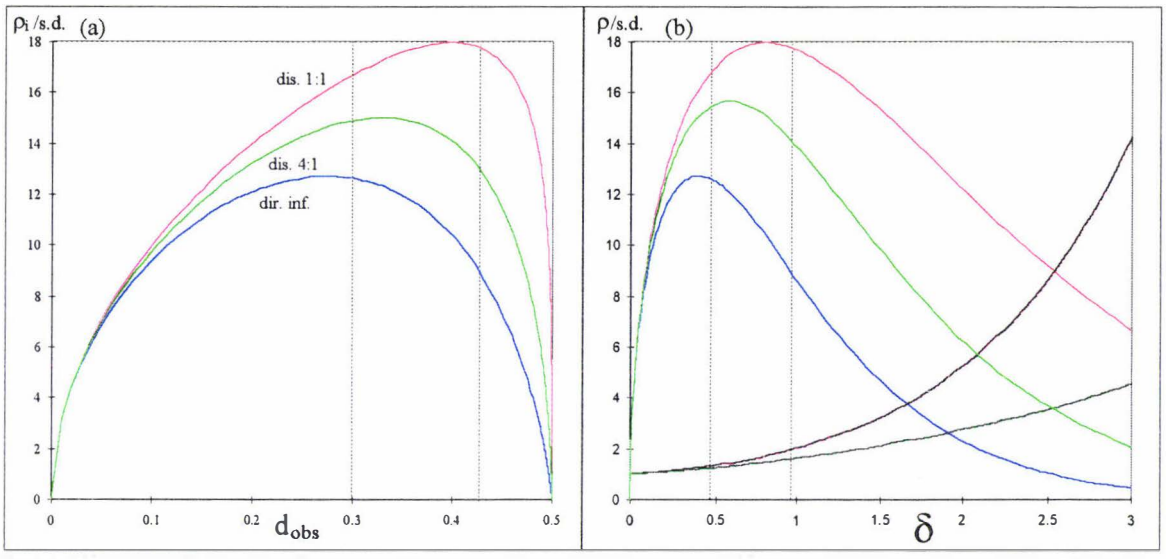


FIGURE **4.10a + b** The signal-to-noise ratio of direct inference of a 4-tipped pathsets length (blue) vs the signal-to-noise ratio when estimating it as a sum of distances. **4.10a** The x-axis is the observed pathset length (measured across all 4 tips simultaneously), while in figure **4.10b** it is the true pathset length (δ) measured in the expected number of substitutions per site. If the pathset is estimated from pairs of pairwise distances which are equal, or 1:1, in size, the signal-to-noise ratio of this estimate is shown as the red line. If the relative size of the distances making up the two "halves" of the 4-tipped pathset are quite unequal in size (specifically in a 4:1 ratio), the signal to noise ratio is indicated as the green line. The first vertical dotted line to the left in each plot marks the pathset length used in our model tree (see table 4.7), while the next vertical dotted line marks the point where the sum of distances method (with both distances equal, red) has a signal-to-noise ratio twice as good as when using the direct pathset length estimate. In **4.10b** only, the dark red (for 1:1 distances) and the dark green (for 4:1 distances) lines in the figure on the right, plot the ratio of the accuracy of the distance methods vs the direct inference equation.

4.7.2 A comparison of the structure in $\mathbf{V}[\hat{\gamma}_D]$ vs $\mathbf{V}[\hat{\gamma}]$

We have calculated the entries in $\mathbf{V}'[\hat{\gamma}_D]$ from $\mathbf{V}'[\hat{\rho}_D]$ (see table 4.7) and now interpret how $\mathbf{V}[\hat{\gamma}_D]$ and $\mathbf{V}[\hat{\gamma}]$ differ. For our model data, figure 4.11 shows the entries in the covariance matrix of $\hat{\gamma}$ from distances, $\mathbf{V}[\hat{\gamma}_D]$, plotted against the corresponding entries in the covariance matrix from sequences, $\mathbf{V}[\hat{\gamma}]$. It can be seen that most of the variances of entries in $\hat{\gamma}_D$ decreased relative to those in $\hat{\gamma}$ ($\hat{\gamma}_1$ and $\hat{\gamma}_4$ from 0.434 to 0.369, $\hat{\gamma}_5$ from 0.103 to 0.046 and $\hat{\gamma}_6$ from 0.025 to 0.016), two remained equal ($\hat{\gamma}_2$, $\hat{\gamma}_7$ at 0.06) while the last ($\hat{\gamma}_3$ relating to the trees internal edge) increased slightly from 0.068 to 0.076. The most important change from the point of view of tree selection was that the variance of the potentially misleading "long edges attract" signal $\hat{\gamma}_5$ has been more than halved, and more than makes up for a slight increase in the sampling variance of $\hat{\gamma}_3$.

Figure 4.11b shows a plot of the correlations of entries in $\hat{\gamma}_D$ vs those in $\hat{\gamma}$, and it is clear there are some differences here (otherwise all entries would lie on the line y = x). Overall correlations have increased in $\hat{\gamma}_D$ (hinted at by the general increase of the small covariances, shown fig. 4.11a). Most importantly for tree selection, the correlation between $\hat{\gamma}_3$ and $\hat{\gamma}_5$ has

gone from being nearly zero to approximately -0.4. This means more sampling overlap between these two variables, than if the correlation was zero. So overall, while there has been an important and substantial reduction in the variance of the potentially misleading entry $\hat{\gamma}_5$, some of this advantage for tree selection has be negated by the negative correlation between $\hat{\gamma}_3$ and $\hat{\gamma}_5$, and the slight increase in the variance of $\hat{\gamma}_3$.

Simulations were run to assess the accuracy of $\mathbf{V}'[\hat{\gamma}_D]$, given that it was using two approximations, namely the minimality criteria (rather than knowing which pair are really disjoint) and the delta method. The results (not shown) indicated that all entries in both the covariance and correlation matrices were accurately predicted when c was 1000 (all entries agreeing to 1% or better). When c was dropped to just 100, a slight underestimate (approximately 5%) of the variances became apparent, while some correlations in the simulation were slightly different from inferred in $\mathbf{C}'[\hat{\gamma}_D]$. Most notably, the near total negative correlation ($\approx$ -1) of 5 and 6 was reduced to about -0.75. This difference is probably due to sampling error causing pairs of non-disjoint distances to be minimal (and so they are used to infer the length of the 4-tipped pathset). This randomisation factor can in turn be predicted to break down some of the strong negative correlation of $\hat{\gamma}_5$ and $\hat{\gamma}_6$. Importantly the correlations of entries $\hat{\gamma}_3$ and $\hat{\gamma}_5$, plus $\hat{\gamma}_3$ and $\hat{\gamma}_6$ where accurately estimated in $\mathbf{C}'[\hat{\gamma}_D]$, even with this short sequence length.

If it were deemed essential to avoid the downward bias inherent in the minimality criterion for choosing the value of a pathset from pairwise distances, we might consider taking some weighted average of all the alternative sums of distances. An appropriate weighting would be according to the probability, that by chance, one of these other pathsets might be smaller than the minimal one chosen. However such a method could reduce the efficiency of tree selection when the internal edge(s) separating the two halves of a quartet were small. The added calculations that such weighted pathset evaluations require (involving not only variances, but also covariances, and especially as the number of taxa increases) make this refinement at present seem overly complicated, but may be worth consideration if the distance Hadamard turns out to be an especially useful method in phylogenetic research.

(figure next)

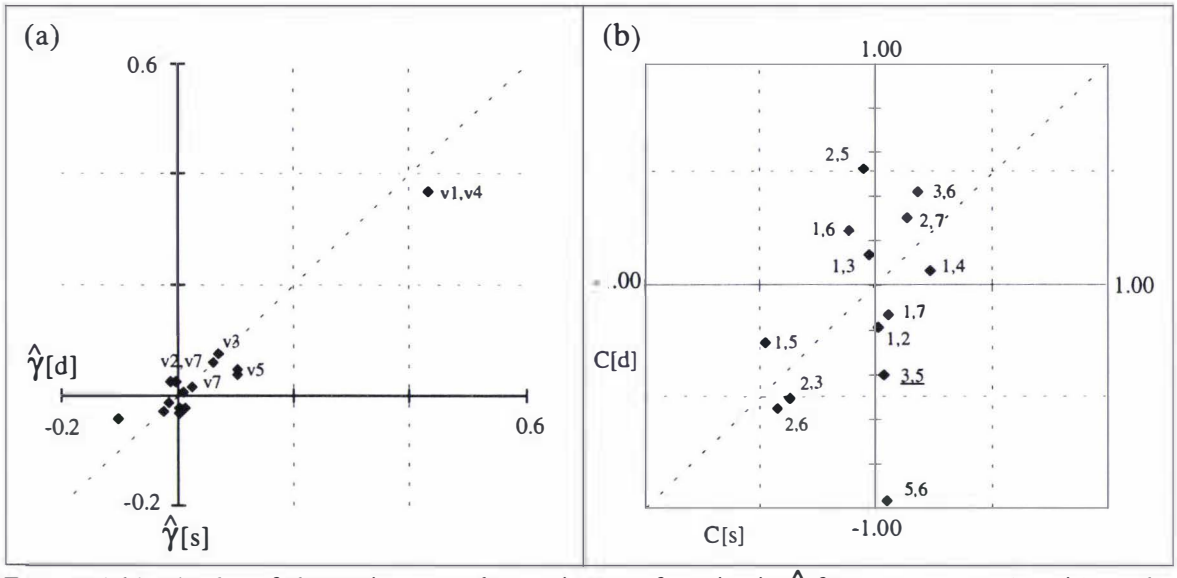


FIGURE **4.11a** A plot of the variances and covariances of entries in $\hat{\gamma}$ from sequences (x-axis) vs the corresponding values in $\hat{\gamma}_D$ (y-axis), estimated from the data in table 4.7 (this used the delta method approximation applied to our four taxon illustrative model of section 4.2.1, with $\lambda = 1$). The numbers have yet to be divided by a specific sequence length (i.e. they use a nominal sequence of length 1). In general the variances of $\hat{\gamma}_D$ (labeled v1-v7 part a) are lower than those in $\hat{\gamma}$. Many of the covariances are also larger in $\hat{\gamma}$. **4.11b** The correlations of $\hat{\gamma}_D$ (y-axis marked C[d]) plotted against those of $\hat{\gamma}$ (recall the correlation of (x, y) is the cov(x,y) / (s.d.(x) × s.d.(y)). The most important correlations for tree selection methods are those between the three signals which may be selected as an internal edge (3, 5, 6). A real concern is that in a strong negative correlation has appeared between the signal for the tree's internal edge (3) and the long edges attract signal (5) (this entry which is underlined has near zero correlation when the full Hadamard conjugation is used). This negative correlation means that when signal 3 is large then 5 tends to be low, but when 3 is low, 5 is expected to be positive, increasing the probability it will be selected as optimal.

4.7.3 The statistical structure of $\hat{\gamma}_D$ vs $\hat{\gamma}$ evaluated on a six taxon tree.

In this section, we look for differences in the comparative structure of γ_D reconstructed from distances versus γ reconstructed from sequences, using a larger tree of 6 taxa. With this many taxa, pathsets with 4 or more tips now outnumber pairwise distances in **r** and ρ, i.e. under the 2-state model, there are 15 pairwise distances, 15 'quartets', and one 'sextet'. In order to keep the complexity of interactions down we use a tree which is well resolved, except for a short innermost edge. This weighted tree (and scaled versions of it) are shown in figure 4.12. This type of tree is often encountered in real data sets; it can be viewed as a 'Felsenstein zone' tree where a biologist has added in two sequences which intersect the longest edges and serve to reduce the long edges attract effect (a good example would be adding a deeply diverging eubacteria and eukaryote to the data used in figure 2.7). For this section we have called the trees by the number of changes on their external edges, having already fixed the sequence length to be studied at 1000. Later in chapter 5, where we will study tree selection with them, they are referred to simply by the expected number of changes per site on each pendant edge e.g. the 160 tree =

160/1000 changes per site on the long edges = 0.16 substitutions expected per site. The model used for these simulations is described further in the caption to figure 4.12. In this section, there is a substitution of notation, $\hat{\gamma}(s)$ for $\hat{\gamma}$, and $\hat{\gamma}(d)$ for $\hat{\gamma}_D$, in order to make it clearer which type of vector is being referred to (since the discussion will be switching from one to the other frequently, and using subscripts to indicate specific entries).

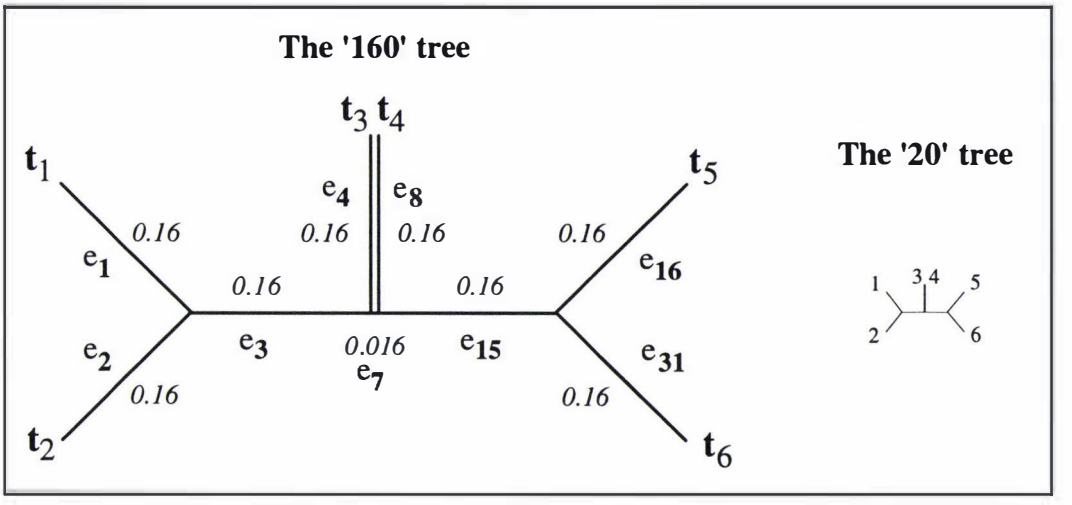


FIGURE 4.12 The six taxon trees used to study further the differences in the statistical structure of $\hat{\gamma}_D$ relative to $\hat{\gamma}$ derived via the Hadamard conjugation. On the larger tree the edge weights are given in italic, taxa numbers are in plain text, while edge indices (by the 2-state Hadamard conjugation notation) are in bold. The sequence length for the study was 1000 sites, and all but one edge on the '160' tree had an expected length of 0.16 × 1000 = 160 substitutions. The short internal edge was expected to incur 16 substitutions. The '80' tree (not shown) had the all edge weights divided by 2, while the '20' tree, has all edge weights divided by 8 (relative to the 160 tree). The basic tree structure is what may occur when one tries to improve phylogenetic inference on a tree like that of Felsenstein (1978a), by adding extra taxa to reduce the long edges attract problem (Felsenstein 1978a, Hendy and Penny 1989). The '160' tree is similar to that inferred for some rRNA and protein sequences used to study the relationships of archaebacteria, with the edges 1 and 2 for example representing divergent eukaryotes (e.g. a vertebrate and *Giardia* sp.), 5 and 6 representing diverse eubacteria (e.g. *E. coli* and *Thermotoga* sp.), while 3 and 4 represent say, an extreme thermophile and a *Halobacterium* (by this labeling we do not mean to infer that the eocyte tree is correct, but rather, wish study the sorts of problems that might be expected to occur should it be the correct tree). In this tree, the gamma vector entry 7 (1+2+4) corresponds to the short internal edge, while gamma 11 (1+2+8) and gamma 12 (4+8, i.e. taxa 3 and 4 being drawn together) correspond to the only likely alternatives to this edge. This is because γ_{11} and γ_{12} are the only signals compatible with the large internal edges 3 and 15 which are chosen well over 99.99 % of the time (chapter 5 tree selection simulations). The model used to predict the $\mathbf{s}(T)$ vector was the 2-state Poisson model.

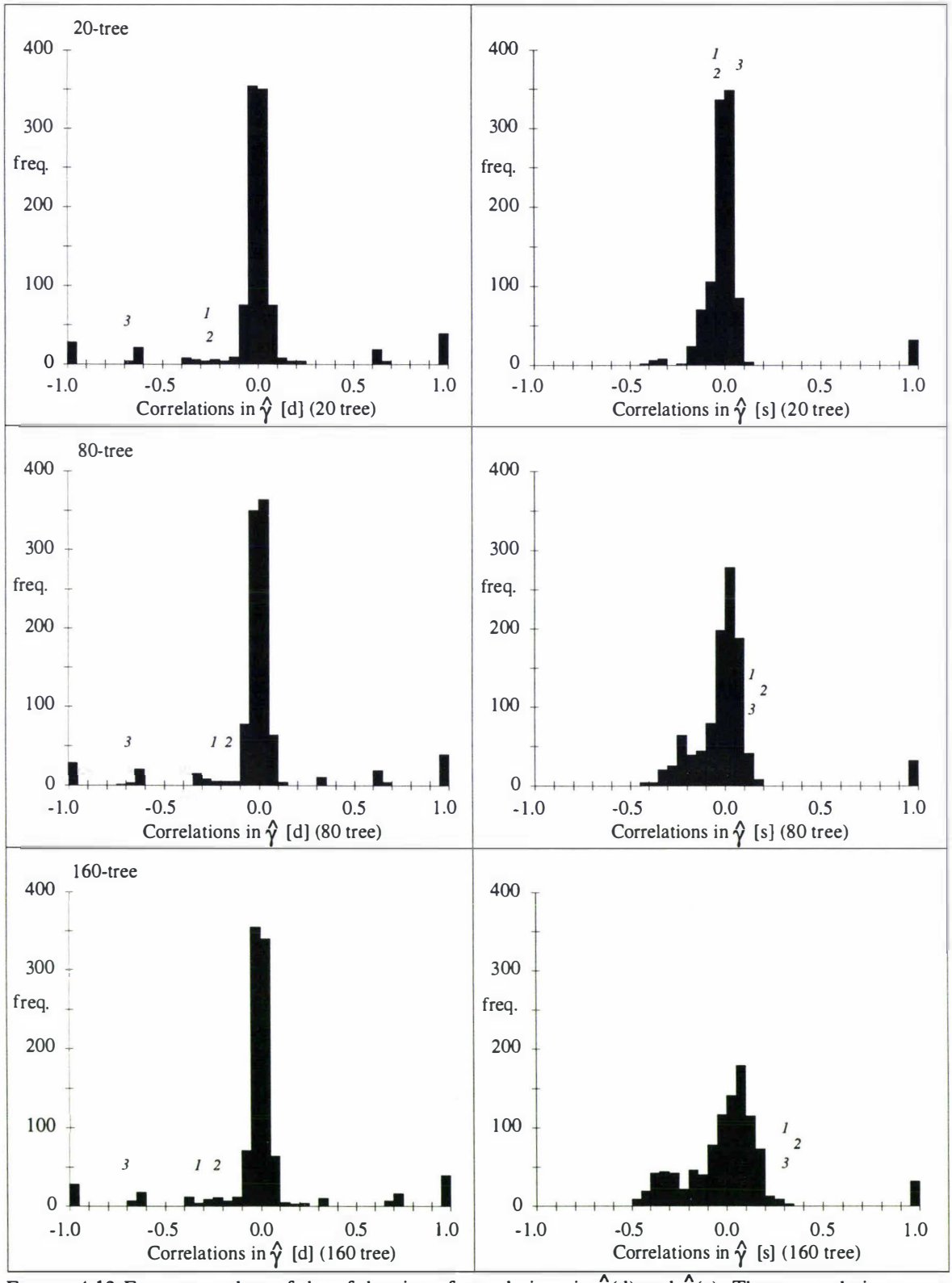


FIGURE 4.13 Frequency plots of the of the size of correlations in $\hat{\gamma}(d)$ and $\hat{\gamma}(s)$. These correlations were estimated from 1,000 random samples of $\hat{\gamma}$ under the models described in figure 4.12 (c = 1,000 in all cases). The top two figures are for sequences derived from the '20' tree, the middle two from the '80' tree, and the last two from the '160' tree. The correlations of $\hat{\gamma}_7$ with $\hat{\gamma}_{12}$ are marked as *1*, $\hat{\gamma}_7$ with $\hat{\gamma}_{11}$ is marked *2*, while that of $\hat{\gamma}_{11}$ with $\hat{\gamma}_{12}$ is marked *3* in each case. The index assigned to these correlations, is in the same rank as their importance in deciding how often an incorrect tree is selected. The more positive the correlations marked *1*, *2* and *3*, the higher the probability of recovering the true tree (and the opposite for negative correlations).

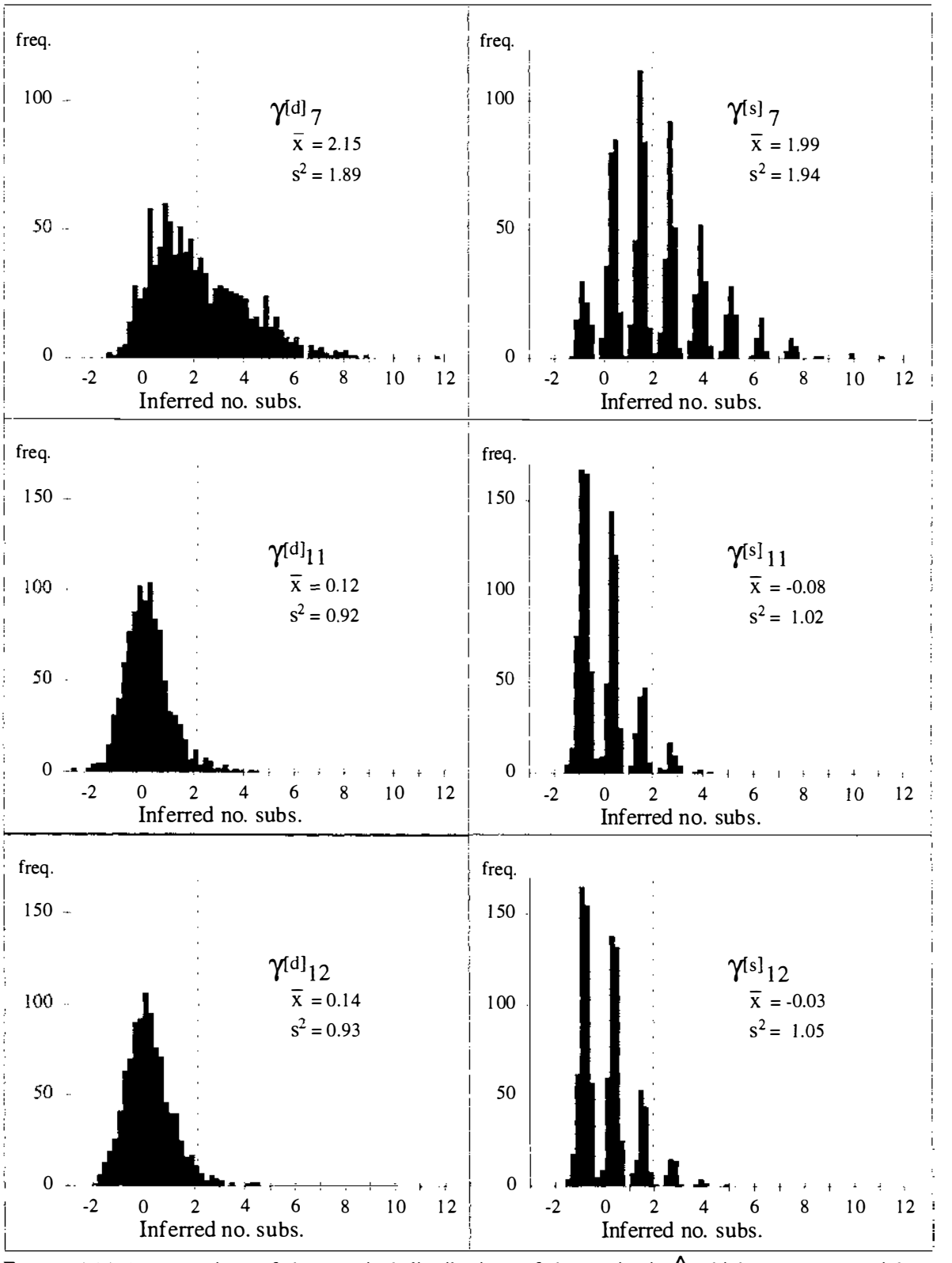


FIGURE 4.14 A comparison of the marginal distributions of the entries in $\hat{\gamma}$ which are most crucial to correctly resolving the '20' tree shown in figure 4.12. The entries in $\hat{\gamma}(d)$, in contrast to those of $\hat{\gamma}(s)$, are essentially continuous. In addition the entries of $\hat{\gamma}(d)$ have slightly smaller variances. We note a slight positive bias to the means (marked with vertical dotted lines) of these three marginal entries in $\hat{\gamma}(d)$. This bias of the means, is larger than expected due to sampling error alone (i.e. s.d.(mean of x) = s.d.(x)/$\sqrt{n}$ (n = 1,000 simulations), so the s.d. (mean of $\hat{\gamma}(d)_7$) $\approx$ 1.9/31.6 = 0.06, while s.d.(mean of $\hat{\gamma}(d)_{11}$ or $\hat{\gamma}(d)_{12}$) $\approx$ 0.92/31.6 = 0.03). This positive bias runs counter to that of the multinomial like entries of $\hat{\gamma}(s)$, which appears, if anything, to have a slight negative bias. Thus, even for this low amount of sequence divergence

(the longest path through the tree is just 0.08 subs. per site), the variances of the crucial entries in $\hat{\gamma}(d)$ are slightly lower, the correlations are slightly less favourable, and the bias is all but insignificant. The performance of standard tree selection algorithms on these two types of spectra should be very similar for this low rate of substitution. Even at this low rate of substitution, the marginal distributions of $\hat{\gamma}(d)_{11}$ and $\hat{\gamma}(d)_{12}$ are near normal, but surprisingly that of $\hat{\gamma}(d)_7$ is substantially skewed (the marginals of are similar to binomial, as expected from the results of section 4.3).

The first noticeable feature of the covariance matrix of $\hat{\gamma}(d)$ ($\mathbf{V}[\hat{\gamma}(d)]$) in our six taxon example (matrix not shown), is that almost all of the variances are less than their corresponding values in $\hat{\gamma}(s)$. In general, these variances have decreased by between 2 and 20 %. Of particular importance to tree selection algorithms applied to these spectra, are the variances of the three entries in $\hat{\gamma}$ (namely 7, 11 and 12) which are compatible with the well supported internal edges 3 and 15 of our model tree. Figure 4.14, shows that the lower variances of entries in $\hat{\gamma}(d)$ (relative to those in $\hat{\gamma}(s)$) are detectable even at low amounts of sequence divergence (in this case they are assessed on the "20" tree, where the longest path is just 0.08 substitutions per site).

The covariances (and also correlations) between entries in $\hat{\gamma}(d)$ are quite different to those in $\hat{\gamma}(s)$. This feature was investigated further by looking at how the correlations of entries within both $\hat{\gamma}(d)$ and $\hat{\gamma}(s)$, changed as the overall amount of divergence increased. The results of this study are shown in figure 4.13. Notice how the correlations of entries within $\hat{\gamma}(d)$ are diverse even at very low rates of change (i.e. in the "20" tree some entries have very high negative correlations), yet change very little as the rates of change increase (the "80", and "160" trees). In contrast the correlations between entries in $\hat{\gamma}(s)$ while all very similar (and near zero) for low rates of change, spread out as the rate of change increases (with a skew tail going towards more negative correlations).

Of particular importance to tree selection are the correlations of the three entries in gamma which are able to be chosen to resolve the short internal edge of our model trees. We marked these correlations as *1* (corr.[$\hat{\gamma}_7$, $\hat{\gamma}_{12}$]), *2* (corr.[$\hat{\gamma}_7$, $\hat{\gamma}_{12}$]), and *3* (corr.[$\hat{\gamma}_7$, $\hat{\gamma}_{12}$]) for our model simulation data, shown in figure 4.13. In $\hat{\gamma}(d)$ *1*, *2*, and *3* are all moderate sized negative correlations. In contrast *1*, *2*, and *3* for $\hat{\gamma}(s)$ have correlations that are near zero for low rates of change and become increasingly positive as the amount of divergence increases. Negative correlations between $\hat{\gamma}_7$ (corresponding to an internal edge in the generating tree) and either $\hat{\gamma}_{11}$ or $\hat{\gamma}_{12}$ (long edges attract) are not desirable since they mean that when $\hat{\gamma}_7$ takes on a small value (due to sampling), there is a better than even chance that either $\hat{\gamma}_{11}$ or $\hat{\gamma}_{12}$ will be bigger than its mean value (making erroneous tree selection more likely). Positive correlations are in contrast desirable, having the opposite effect. The negative correlation between $\hat{\gamma}_{11}$ and $\hat{\gamma}_{12}$ of $\hat{\gamma}(d)$ is also undesirable, since it indicates that when $\hat{\gamma}_7$ takes small values, either one or the other of the two entries competing with it for selection will tend to take on a larger than usual value (hence approximately doubling up the effect of the aforementioned negative correlations). Since all these correlations are of moderate size, they seem unlikely to completely offset the large

decrease in the variances of the crucial three entries in $\hat{\gamma}(d)$. This expectation is shown to hold true in the tree selection simulations of chapter 5.

The marginal distributions of $\hat{\gamma}(d)$, compared to those of $\hat{\gamma}(s)$, are quite distinct with short sequences, or corresponding to entries in $\hat{s}$ with low expected values.. While only the marginal distributions of $\hat{\gamma}_7$, $\hat{\gamma}_{11}$ and $\hat{\gamma}_{12}$ are shown in figure 4.14, our comments generally hold for the marginal distributions of the other entries in $\hat{\gamma}$ (results not shown). In figure 4.14 it can be seen that the variances of the two non-tree entries have dropped by approximately 10% (in $\hat{\gamma}(d)$ relative to $\hat{\gamma}(s)$), while that of the internal edge has dropped by only 2% or so. More strikingly, however, the marginal distributions of entries $\hat{\gamma}(d)_{11}$ and $\hat{\gamma}(d)_{12}$ are very similar to normal distributions (with only slight tails to the right), while their counterparts in $\hat{\gamma}(s)$ are clearly binomial. This means that even with small amounts of sequence divergence, statistical tests on entries some entries in $\hat{\gamma}(d)$ can be made using a normal approximation and the diagonal entries of $\mathbf{V}'[\hat{\gamma}(d)]$ (however we have yet to assess the sampling properties of $\hat{\mathbf{V}}'[\hat{\gamma}(d)]$ as exhaustively as we did for $\hat{\mathbf{V}}'[\hat{\gamma}(s)]$).

The strong skewness of $\hat{\gamma}(d)_7$ (figure 4.14) shows that it cannot always be assumed that entries in $\hat{\gamma}(d)$ have near normal marginal distributions (nor that the worst violations of this assumption, coincide with the smallest entries). The skewness of this entry is probably due, in part, to a cumulative bias introduced by the minimality criteria in estimating higher order pathset lengths as sums of distances, combined with their own binomial marginal distributions. Further, it is conjectured that the marginal distribution of the entries in $\hat{\gamma}(d)$ corresponding to a short internal tree edge are most affected in this way. So, unfortunately, it can be expected to occur in just those situations where we most desire accuracy of our tests (i.e. about poorly resolved internal edges). If a normal approximation is to be used, it should be realised that the corresponding test for such an entry being zero, will be rejected too often. Fortunately, this skewness diminishes quite quickly as the overall rate of change increases, or as sequence length increases. A conservative rule for assuming normality of the marginal distributions in $\hat{\gamma}(d)$ would thus appear to be: assume normality only if the variance of entry $\hat{\gamma}_i$ is $\geq (5/c)\times([c-5]/c)/c \approx (5)/c^2$ or the s.d. of $\hat{\gamma}_i$ is $\geq 2.24/c$ (where c is the sequence length).

Another feature apparent from figure 4.14 is that there is a slight bias on the means of entries of both $\hat{\gamma}(d)$ and $\hat{\gamma}(s)$. The means for $\hat{\gamma}(d)$ are all slightly overestimated (by about 0.14 substitutions), whereas those of $\hat{\gamma}(s)$ are even more subtly underestimated. However, this usually will not present a serious problem, and this bias decreases quite rapidly with increasing sequence length.

Apart from the trees discussed so far, I have examined the marginal entries of many other spectra generated for many combinations of tree size (up to 8 taxa), shape, sequence lengths, and overall rates of change. Nothing that I observed contradicted any of the generalisations made so far. The most notable feature of any spectra, not already mentioned, is what I call the "spider".

The "spider" is a slightly unusual marginal distribution observed on just a few entries in $\hat{\gamma}(d)$ only. It is a highly symmetric distribution (very near the normal in outline) which showed very discrete classes, like a binomial distribution with a mean of greater than 5. It is expected that this was caused by a particular weighted tree, such that the entries in $\hat{\gamma}(d)$ were the results of adding and subtracting many similar sized distances, and preserving something of the discrete sampling distribution of the observed distances. It is a curiosity, which will be studied further at a later date. The symmetry of these entries, their approximately normal shape and their rarity suggest that this type of marginal distribution poses no particular problem for either tree selection, or hypothesis testing.

4.7.4 Variations on the distance Hadamard

An exception to sequences yielding just one distance matrix, are models where the rates of transitions and the two types of transversions can vary independently, but still allow the recovery of a tree additive distance matrix of each type of change. For 4 states the most general of these models is the generalised Kimura 3ST model (Evans and Speed 1993). Under this model it is possible to derive three completely separate distance matrixes for transitions, transversions type 1 and transversions type 2 respectively. Each one may then independently yield a distance spectra, which can used and studied in the same way as the equivalent elements of the 4^{t-1} Hadamard conjugation or the order 2t-1 4-state conjugations (see appendix 2.6). These separations of transitional and transversional changes can be used to estimate the tr / tv ratio after making corrections, but without needing to estimate a tree. They also allow separate examinations of how much phylogenetic information each type of change is retaining. It will be interesting to explore the effect of weighting each of these distance matrices (in a manner analogous to that discussed in section 4.5), prior to tree selection from up to three separate $\hat{\gamma}(d)$ vectors.

Of course, under the generalised time reversible model (and its extension to allow rates across sites in section 3.2.2), the whole rate matrix is estimated (as $\mathbf{R}t = M^{-1}(\mathbf{P}) = M^{-1}(\Pi^{-1}\mathbf{F})$), allowing up to 12 corrected types of substitution to be estimated separately (which are the product of three independent equilibrium base frequencies, in which case $\mathbf{R}t$ has up to six more independent parameters). That is, the corrected number of A→ C changes will be $\pi_a(\mathbf{R}t)_{12}$, A → G = $\pi_a(\mathbf{R}t)_{13}$, ... , T → G = $\pi_t(\mathbf{R}t)_{43}$ (e.g. see Lanave *et al.* 1984, Tavaré 1986, and sections 3.2.1 and 3.2.2). A similar division into types of changes can be done with many amino acid distances. These procedures are strictly valid only under the general time reversible model's assumptions, and unlike the Kimura 3ST and submodels, do not allow rates to change independently on different edges in the tree.

A simpler (but still consistent) approach to examining transitions separately to transversions, is constructing a neighbour joining tree for each distance matrix and then comparing edge weights (either as a sum if using the general time reversible distances, or on particular edges for the generalised Kimura 3ST and 2ST). Other methods such as split decomposition (Bandelt and Dress 1992), or weighted least squares can also be used to study each type of change separately.

Of course the tree based methods, will not show up anomalies as will the distance Hadamard or split decomposition. Most simply the overall magnitude of entries in these three matrices can be compared in order to assess things such as transition to transversion ratio's (or a molecular clock), without needing to estimate a tree. These separated substitution distance Hadamard spectra will have entries in γ with even lower variances than those in the corresponding constrained order 4^{t-1} Hadamard conjugations (section 4.6), or the order 2^{t-1} Kimura 3ST Hadamard conjugations (appendix 2.6). The price of reduced variance, appears to be less predictable behaviour with regard to specific model violations, and perhaps less reliability for tree selection when the model's assumptions are violated. It is the former property which compromises a biologist's ability to diagnose potential trouble with phylogenetic analysis, which may be the more regrettable.

Another potentially useful variation suggested by this study, is the estimation of only some of the higher order pathsets by sums of lower order pathsets. A simple example would be to use all pairwise distances plus quartet entries in ρ, then infer all higher order pathsets from a combination of these values using the logical extension of Buneman's 4 point condition. A more refined approach might be to replace just those pathsets with variances larger than those of the largest distance. Such hybrid methods (of which there are many possibilities) would aim to reduce the variance of entries in $\hat{\gamma}$, yet not throw away much of the information that allows accurate diagnosis of the cause of model violations. Much study could be done in this area if it Hadamard conjugations do turn out to have a long term place in the study of sequence evolution. Such modifications still allow model exact modifications to allow for a distribution of rates across sites, as developed in chapter 2.

4.8 THE MEANING OF $\hat{\gamma}$, AND ESTIMATING TREE SELECTION RELIABILITY

This section considers two important issues that arose during these studies, which are not directly related to each other. The first regards the theoretical statistical nature of the vector $\hat{\gamma}$, while the second deals with the practical question of using knowledge of the sampling distribution and covariance matrix of $\hat{\gamma}$ to infer the probabilities of picking different trees in a simulation, or make an estimate of these probabilities from sampled data. It also considers how non-linear transforms bias bootstrap approaches to estimating variance. This bias, is in addition to the biases in tree selection identified by Zharkikh and Li (1992a, 1992b).

4.8.1 Are Hadamard conjugations ML estimators?

Here we consider whether entries in $\hat{\gamma}$ are ML estimators of a particular kind (under the 2-state Poisson model). Recall that the vector $\hat{\mathbf{s}}$ is, under an i.i.d. model, the ML estimator of the vector $\mathbf{s}(T)$. The multiplication of $\hat{\mathbf{s}}$ by $\mathbf{H}$, is an orthogonal unweighted linear transformation.

The entries in $\hat{\mathbf{r}}$ that correspond to single paths, are ML estimates of $(1\text{-}2d_{obs})$, and so the ML estimate of $d_{obs}(i) = -(1-\hat{r}_i)/2$. Thus $\hat{\mathbf{r}}$ is a vector of ML estimates of 1-{twice the observed pathset lengths}.

The logarithmic transformation to make pathset lengths additive (in expectation), is also an ML estimator. It is analogous to the 4-state Poisson model transformation (i.e. the Jukes-Cantor equation), which for some time has been recognised as an ML estimator (e.g. Saitou 1990). For this discussion, we call entries in $\hat{\rho}$ ML estimates, recognising, of course, the multiplicative constant -2. This constant can be negated at will, and would otherwise be needed to "balance" the multiplication by **H** (for example, let **x** be the vector of ML pathset length estimates = $-1/2\hat{\rho}$, then $\hat{\gamma} = -2\mathbf{H}\mathbf{x}$). Some entries in $\hat{\rho}$ correspond to ML estimates of total pairwise distances (including multiple substitutions at a site), while the remainder are ML estimates of the length of higher order pathsets (except $\hat{\rho}_0$ which invariably estimates the length of the null pathset i.e. 0). This interpretation follows because Hendy (1989) shows that in order to estimate the length of a pathset, it is only necessary to know the observed pathset 'length' then apply the logarithmic transformation (we have already established that the ML estimates of the observed pathset lengths are entries in $\hat{\mathbf{r}}$ which have binomial marginal distributions). Just like the entries in a matrix of ML pairwise distances (of which entries in $\hat{\rho}$ form a superset), all the entries in $\hat{\rho}$ are estimated independently of needing to specify any particular tree or set of trees.

The vector $\hat{\gamma}$ is an orthogonal unweighted linear transformation of $\hat{\rho}$. Asymptotically (as $c \to \infty$, and given that the Poisson tree model already described holds) the entries in $\hat{\gamma}$ converge to either zero (if they are not in the true tree) or else the positive length of an edge in the true tree. Recognising this, we suggest that the entries in $\hat{\gamma}$ are ML estimates of the corresponding entries in $\gamma(T)$, which are inferred with no conditioning on a specific tree or subset of trees. Another way of describing this is: entry $\hat{\gamma}_i$ is the ML estimate of the length of the putative edge e_i, inferred with calculations that do not involve any specific subset of trees (or that the calculations are made taking account of all possible trees equally).

One possible objection to $\hat{\gamma}$ being an ML estimator of $\gamma(T)$, is that entries in $\hat{\gamma}$ nearly always have larger variances than those in $\hat{\gamma}_D$ (as measured under the model), despite the usual proof that ML estimators are asymptotically the most statistically efficient (i.e. have least variance) estimators. The main reason that $\hat{\gamma}_D$ has lower variances on many of its entries is that it infers pathset lengths conditional on using information from a small set of trees, which under the model are all expected to be similar to the true tree. This is obvious, when it is highlighted that the distance Hadamard method, makes an inference of the quartet structure of the true tree (based upon Buneman's 4 point metric) every time it infers the length of a quartet. This 4-point condition is essentially a four taxon tree selection criterion, so effectively, inferring the 4-tipped pathsets in $\hat{\rho}_D$ is a by product of inferring all possible four taxon trees. As Steel *et al.* (1993b) prove, if you know all the quartet relations in a tree, then you can uniquely reconstruct that tree. And of course, the distance Hadamard method also uses the minimum sum of three or more

distances to estimate higher order pathsets and these also strongly condition on a subset of trees. So effectively, the distance Hadamard has reduced the variance of entries in $\hat{\gamma}_D$ by invoking a form of tree selection before inferring entries in $\hat{\gamma}_D$. This is something we specifically claim the Hadamard conjugation does not do. Thus, our claim of $\hat{\gamma}$ being a specific type of ML estimator is not falsified by its entries generally having larger variances than those in $\hat{\gamma}_D$.

Our claim does not address how entries in $\hat{\gamma}$ are related to another well known ML estimator often used in phylogenetics, specifically the ML method of tree selection. However, as we show in chapter 5, the edge length estimates inferred by $\hat{\gamma}$ and the ML tree selection method are expected to differ (even when the data comes form the 2-state Poisson model). This difference is expected, since they are quite different sorts of ML estimators. The ML tree selection method cannot even begin to make an edge length estimate without conditioning explicitly upon a particular tree. It may have been an assumed identity between edge length estimates made by the ML tree selection criterion, and those made in $\hat{\gamma}$, that lead Hendy and Penny (1993) to suggest that choosing a tree using closest tree from $\hat{\gamma}$ was effectively the same as the ML tree selection criterion of Felsenstein (1981a). In chapter 5 we show that no such identity between closest tree and ML exists, yet another method of tree selection applied to $\hat{\gamma}$ does share a close identity with Felsenstein's (1981a) method. Yet it is not to this identity we claim.

An important question is: what use is a particular ML estimator, based on a particular model, when alternative methods or modifications achieve a specific task (e.g. edge length estimation) more efficiently under that same model? It seems an unavoidable conclusion of these studies, that the Hadamard conjugation generally does relatively less well than $\hat{\gamma}_D$, as a starting point for estimating both the unweighted tree and it edge lengths, when the model holds (this prediction is born out by the model of tree selection described immediately below, the simulations in chapter 5, and is consistent with the simulation results of Charleston *et al.* 1994). Rather than this somewhat narrow view of utility, we emphasise results in chapter 5, which suggest that $\hat{\gamma}$ can be most useful, not when all the assumptions of the model hold, but rather when the model is violated in specific ways. For example, even though $\hat{\gamma}$ is not exact when data comes from a mixture of two or more trees, it can reliably extract important information lost by methods which need to condition upon a specific tree. An important example of this is given in section 5.4.

Thus it is conjectured that $\hat{\gamma}$ is a specific type of ML estimator, analogous in some ways to a matrix of ML pairwise distances. It potentially finds its greatest use when the model is violated (ironically, where it may no longer be an ML estimator). A related conjecture, which we will not delve into, is that $\hat{\gamma}$ may be an ML estimator (of some type) of the sum of $\gamma(T')$'s when the process of sequence evolution is 2-state i.r. / i.i.d. Poisson, but the **s** vector analysed is a sum of sequences from different trees (and again it is an ML estimator which does not condition on specific trees). It should not be confused with ML estimators which condition on a specific set of weighted trees (e.g. as in section 5.4). Under these types of model, the data has not been

generated by a single tree, but perhaps by a network of evolution, or even disjoint processes juxtaposed by recombination. It is important to re-emphasise the condition that $\hat{\gamma}$ is an ML estimator when the graph describing evolution is unspecified (e.g. do not specify a single tree or a specific form of network). (We are currently looking at the possibility of proving either or both of these conjectures with Dr Mike Steel).

This study also suggests another way that spectral analysis may be extended. It has always been a wish to extend Hadamard conjugations to more general models, but at the present the most general are those developed in chapter 2. Further, it is seems quite certain from the proofs in Székely *et al.* (1993) that the standard conjugation cannot be extended to be exact under, say, a 12 parameter, 4-state model. An insight gained from developing this section is that perhaps we should not expect extensions of spectral analysis to be methods which do not condition on any tree, but rather a method which conditions on all trees. Tying in with this, chapter 5 presents clear evidence for certain non-tree signals showing up as 'significantly nonzero' edges in far from optimal trees. That is, in these trees these edges have much more support than expected under the model, which is precisely what the Hadamard conjugation seeks to estimate. Our proposal to extend spectral analysis is to estimate the support for all 2^{t-1} edge weights by optimising edge weights on all possible binary trees with a method such as ML, which rakes into account expected multiple substitutions. Alternatively, distance methods could be used although they may not be as efficient. Following this, an entry in $\hat{\gamma}(\text{all trees})_i$ ($i > 0$) is the average weight of edges with this index calculated over all binary trees. It is hoped that this new method, which we will call "all trees (or AT) spectral analysis," will identify non-tree signals more reliably and sensitively than $\hat{\gamma}_D$. (The "distance method" $\hat{\gamma}_D$ may only have enough information to look for locally optimal signals; it may do well, although the poor performance of compatibility and closest tree in the "mixture of two trees simulations" in Charleston (1994) figure 5.24 suggests perhaps not (assuming our own interpretations of these results, as given next, are correct)). It may also be useful to consider estimating $\hat{\gamma}(\text{all trees})$ from partially resolved trees and not just binary trees (elsewhere in the thesis, estimating $\hat{\rho}$ from sums of edge lengths of smaller, e.g. 4, 8, .. taxon, ML trees is also considered). Clearly, further investigations beyond the scope of this thesis are needed to identify the strengths of these different methods of spectral analysis. That is, especially their different abilities to highlight, and at least partially diagnose, signals which do not fit the model (which includes the expectation of one tree generating the data, plus a specific mechanism of sequence evolution).

The predicted utility of $\hat{\gamma}$, under violation of the model, may be consistent with a set of simulations performed by Charleston (1994 figure 5.24) where the sequences of two trees are mixed together, and compatibility and closest tree applied to $\hat{\gamma}$ clearly perform best. However, our preferred hypothesis to explain these simulation results is that the combination of sequence data (either transformed or not, but not reduced to just pairwise distances) followed by a compatibility-like tree selection criterion results in the most reliable identification of edges in the tree contributing most to the data. Thus it is the superior performance of compatibility (over parsimony for example) in this situation which is predicted to be making the significant

difference. This is a slightly different interpretation of when compatibility will do well, which adds to the mixture of informative and highly uninformative sites model considered by Felsenstein (1981b) (see also Kuhner and Felsenstein 1994, while in chapter 5 we show that closest tree, which also did well in Charleston's table 5.24, can be considered a form of weighted compatibility). Unfortunately, the matches of tree selection criterion with data transformation in this part of Charleston (1994) are incomplete, so this interpretation will need further study to confirm or refute (the prediction is that parsimony applied to the distance Hadamard transformed data will not do well, but both compatibility and closest tree applied to the observed sequences will do well in identifying the tree contributing most of the data).

4.8.2 Tree selection probabilities estimated via the sampling distribution of $\hat{\gamma}$

Since many methods of tree selection (e.g. parsimony, compatibility, closest tree) applied to the 2-state four taxon Hadamard conjugation simplify to the selection of the largest of $\hat{\gamma}_3$, $\hat{\gamma}_5$ and $\hat{\gamma}_6$, then it is relatively easy to calculate directly the probabilities of selecting each possible tree. For simplicity, it is easiest to start with the case when a criterion selects the most positive of these three entries, even if it does take on a negative value. Further, we will initially make the simplifying assumption that $\hat{\gamma}$ is distributed multi-variate normally, with zero covariance between entries. Thus the probability of one of these tree selection criteria picking the wrong tree is equal to the probability that $\hat{\gamma}_5$ and / or $\hat{\gamma}_6$ is greater than $\hat{\gamma}_3$. Using the illustrative model of section 4.2 and the covariance matrix of $\hat{\gamma}$ given in table 4.4 (in section 4.2.5), and assuming the sequence length is 500, then: $\hat{\gamma}_3$ has mean 0.025 and variance 0.0684/500: $\hat{\gamma}_5$ has mean 0 and variance 0.1031/500: $\hat{\gamma}_6$ has mean 0 and variance 0.0246/500. Given the assumption of multivariate normality and independence (which seems reasonable given the simulation results of section 4.3), then $\hat{\gamma}_3 - \hat{\gamma}_5$ is normally distributed with mean 0.025, and variance = (0.0684 + 0.1031)/500 giving s.d. = 0.0185 (and likewise for $\hat{\gamma}_3 - \hat{\gamma}_6$).

The next step in determining the reliability of tree selection, is to make integrals of the multivariate normal distribution of $\hat{\gamma}$. The probability that is $\hat{\gamma}_5$ is greater than $\hat{\gamma}_3$ is equal to the probability of an observation from a univariate normal distribution (with mean 0.025, and s.d. = 0.0185) having a value of less than zero, which is equivalent to the integral of the tail of a standardised normal distribution beyond 1.350 standard deviations, equals or 0.0885 (to 4 places). By the same argument the probability that $\hat{\gamma}_6$ is greater than $\hat{\gamma}_3$ is equal to the area under the tail of a normal distribution beyond 1.833 standard deviations from the mean (i.e. z = -1.833), or 0.0334. Thus the overall probability of choosing an incorrect tree is 0.0885 + 0.0334 - (0.0885 × 0.0334) ≈ 0.1190, or 11.9%. The last term in brackets is an approximation of the probability that both $\hat{\gamma}_5$ and $\hat{\gamma}_6$ are greater than $\hat{\gamma}_3$ (later it is shown that an exact estimate requires a numerical integration). Despite this last term being inexact, we can put an approximate upper bound on its size. It must be less than 50% of the probability of $\hat{\gamma}_6$ exceeding $\hat{\gamma}_3$.(since all entries are independent, and $\hat{\gamma}_5$ is the larger of this pair). Making these calculations for a whole range of sequence lengths, gives the results shown in figure 4.14. The accuracy of this method for $c < 200$

is uncertain, since simulations show that the marginal distributions of these three entries can still be quite binomial in character (see section 4.3), and because of the approximation made in estimating how often both $\hat{\gamma}_5$ and $\hat{\gamma}_6$ are greater than $\hat{\gamma}_3$.

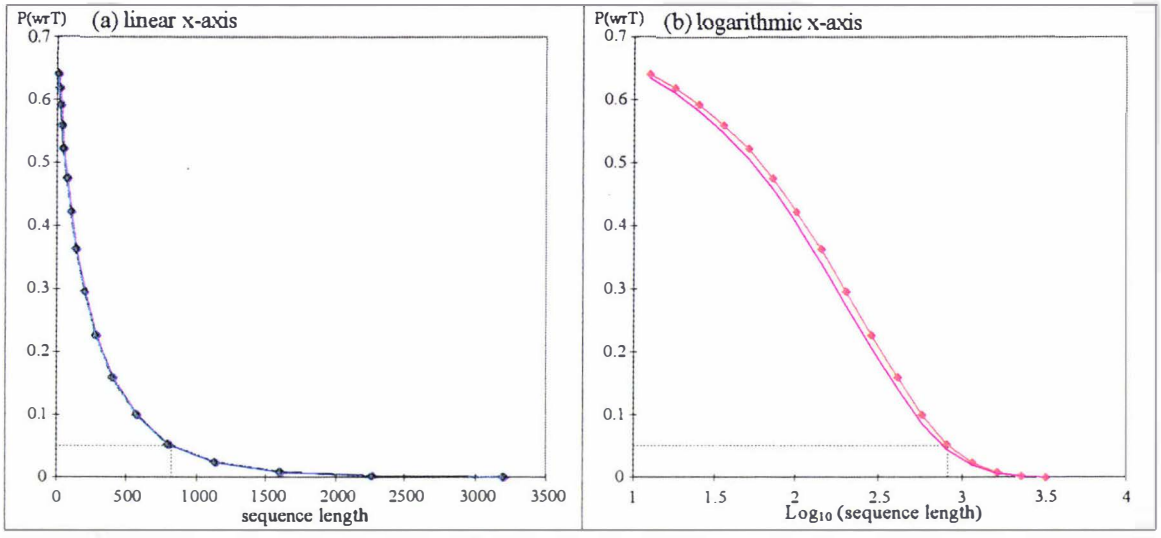


FIGURE **4.15a** The estimated probability, P(wrT), of compatibility, parsimony, or closest tree choosing a wrong tree when applied to $\hat{\gamma}$ for various sequence lengths (with $\hat{\gamma}$ generated by the Hadamard conjugation applied to random samples from the vector s(T) of table 4.1). P(wrT) was calculated according to the method described in section 4.8.2 (assuming independence of the three relevant entries in $\hat{\gamma}$). The expected values and variances of entries in $\hat{\gamma}$ are taken from table 4.4 (in section 4.2.5). **4.15b** As for 4.15a, except the x-axis is now log (base 10) of the sequence length, the red line (with diamonds) refers to the same probabilities as used to generate the blue line in 4.15a, while the purple line batter takes into account the correlations of $\hat{\gamma}_3$, $\hat{\gamma}_5$ and $\hat{\gamma}_6$. The dotted line indicates when the probability of correctly choosing the true tree rises to 0.95 (at this point the sequence length is 827 on the red line and 767 on the purple line). The sequence lengths were generated by the series $i+1 = (i \times \sqrt{2})$ starting with 12.5 and finishing at 3200. The general sigmoidal shape is correct, but the upper shoulder possibly falls away to quickly, because we have not taken into account the non-normal sampling distributions with these short sequences. The sigmoidal shape of these curves is very much like that observed in resampling studies (e.g. Charleston *et al.* 1994), and so, is perhaps best understood as a consequence of converging towards a multivariate normal distribution with increasing sequence length.

The next level of refinement to use in estimating the tree selection reliability, is to take into account the covariances between entries. Using the same example used in the previous paragraph, the variance of $\hat{\gamma}_3 - \hat{\gamma}_5 = \text{var}(\hat{\gamma}_3) + \text{var}(\hat{\gamma}_5) - 2\text{cov}(\hat{\gamma}_3, \hat{\gamma}_5) = (0.0684 + 0.1031 - 2 \times 0.0036)/500 = 0.000329$. Thus if the original data were normally distributed, so too will $\hat{\gamma}_3 - \hat{\gamma}_5$ with mean 0.025, and our refined estimate of its s.d. is 0.0181 (and the probability of $\hat{\gamma}_3 - \hat{\gamma}_5$ being less than zero becomes the standard cumulative normal probability below -1.379 s.d., versus 1.350 previously). The new estimate of the probability $\hat{\gamma}_3 - \hat{\gamma}_5 < 0$ is 0.0839, a slight decrease from 0.0885. When the variance of $\hat{\gamma}_3 - \hat{\gamma}_6$ is re-estimated, the probability of $\hat{\gamma}_3 - \hat{\gamma}_6 < 0$ becomes equal to the cumulative standard normal probability below -2.007 equals 0.0224 (a more substantial change from 0.0334). The covariance of $(\hat{\gamma}_3 - \hat{\gamma}_5)$ with $(\hat{\gamma}_3 - \hat{\gamma}_6)$ is

$-2(\text{var}[\hat{\gamma}_3] + \text{cov}[(\hat{\gamma}_3, \hat{\gamma}_6] + \text{cov}[(\hat{\gamma}_5, \hat{\gamma}_3] + \text{cov}[(\hat{\gamma}_5, \hat{\gamma}_6])$. Thus the probability of choosing the wrong tree is the cumulative density of a bivariate normal with a covariance matrix of the elements specified. The probability of getting an incorrect tree is estimated as the integral of this distribution across all quadrants of the Cartesian plane except the positive quadrant. The other quadrants correspond to one, or the other, or both incorrect trees being more optimal than the true tree. This integral can easily be evaluated with numerical methods. To restrict tree selection to situations in which an optimal tree is only picked if the weight for its internal edge is positive, then it is necessary to do a three dimensional integration about the original three entries $\hat{\gamma}_3$, $\hat{\gamma}_5$ and $\hat{\gamma}_6$.

Similar integrations can be used to estimate the efficiency of tree selection from the distance Hadamard, or standard distance tree selection procedures (e.g. ordinary least squares, neighbor joining, ST method) applied to four taxa. In both the latter cases, correlations can become substantial and should be taken into account as described in the previous paragraph. We have already described how to estimate $\mathbf{V}[\hat{\gamma}_D]$ in section 4.7.1. Since the optimal tree by the previous criteria (i.e. neighbor joining = ST = distance Hadamard with four taxa) is identified by the minimum sum of pairwise distances e.g. min $\delta_{ij} + \delta_{kl}$, then in our example data from s(T) (table 4.1) we want to know how often $(\delta_{13} + \delta_{24}) - (\delta_{12} + \delta_{34}) < 0$, and / or $(\delta_{14} + \delta_{23}) - (\delta_{12} + \delta_{34}) < 0$ (we will call the first quantity A, and the second one B). So the variance of $A = \text{var}(\delta_{13}) + \text{var}(\delta_{24}) + \text{var}(\delta_{12}) + \text{var}(\delta_{34}) - 2\text{cov}(\delta_{13}, \delta_{24}) - 2\text{cov}(\delta_{13}, \delta_{12}) - 2\text{cov}(\delta_{13}, \delta_{34}) - 2\text{cov}(\delta_{24}, \delta_{12}) - 2\text{cov}(\delta_{24}, \delta_{34}) - 2\text{cov}(\delta_{12}, \delta_{34})$ (we obtain the variance of B in the same way). The covariance of A and B is $2\text{cov}(\delta_{13}, \delta_{14}) + 2\text{cov}(\delta_{13}, \delta_{23}) + \ldots + 2\text{cov}(\delta_{34}, \delta_{12}) + 2\text{cov}(\delta_{34}, \delta_{34})$ (the last term being $2\text{var}(\delta_{34})$). Thus, to estimate the probability of any tree other than T_{12} being optimal requires the integral of a bivariate normal distribution with the stated means of A and B, and their covariance, over all sections of the Cartesian plane, except the positive quadrant. With just four taxa, both the distance Hadamard, and the distance tree selection method tend to have marginal distributions which more quickly approach normal than $\hat{\gamma}$ if the amount of divergence is moderate, so this approximation should be more accurate.

It is easy to conceptually extend this method to more than four taxa, although the computations become more expensive, and the validity of the approximations less certain (especially regarding the multi-variate normality assumption). There are, however, many possible simplifying situations, for example when certain edges are almost guaranteed of being picked in the optimal tree (or conversely excluding edges which are very unlikely to be selected). A good example of this, is the model described in figure 4.12. In such cases, any realistic possibility of optimality tends to limited to a few trees (in this example generally three trees), and this cuts down dramatically the number of entries in $\hat{\gamma}$ that need to be considered in order to obtain an accurate approximation.

While these methods make approximations, they do have one useful characteristic in that they can utilize the nearly unbiased estimates of the variances and covariances obtained by the delta method applied to samples. In contrast the bootstrap will overestimate the variances of entries in when applied to samples (due to the effect of the non-linear transformations involved,

as discussed earlier). This would be particularly important when there are large pathset lengths involved. It would also sometimes get around the problem of what to do about estimating $\hat{\gamma}$ when there are "infinite" pathset lengths implied in bootstrap samples, but not in the original data. An approximate estimate of how much bias towards overestimating the variance in entries in $\hat{\gamma}$ there will be due to using the bootstrap on samples, can be obtained by comparing the bias of the delta method, applied to s(T) verses exact sampling (which is asymptotically the same as the bootstrap), as was done in section 2.4. That is, a lower bound of the bias of the bootstrap when estimating variances, which is $\hat{\mathbf{V}}^{bs}[\hat{\gamma}] - \hat{\mathbf{V}}'[\hat{\gamma},]$ will be obtained as $\mathbf{V}[\hat{\gamma}] - \mathbf{V}'[\hat{\gamma}]$. (The logic here being, since the delta method is very nearly unbiased on samples, the bias of the delta method to resampling from s(T) will be of a lower order than the bias of the delta method relative to resampling (the bootstrap) when both are applied to samples of s(T), that is $\hat{s}$. This should hold as a lower bound, because the degree of non-linearity is worse in the second case, as the samples will show larger variances about their true value).

Given an estimate of the bias in bootstrap estimates of the variance of $\hat{\gamma}$, it is possible to then infer the bias of the bootstrap in estimating the reliability of tree selection. In section 4.4.1, we evaluated the bias of $\mathbf{V}[\hat{\gamma}] - \mathbf{V}'[\hat{\gamma}]$ and it was found to be of the order of 15% for the highest rate of change (λ_1, which is being used in figure 4.15, while the overestimation on the smaller entries, $\hat{\gamma}_3$, $\hat{\gamma}_5$ and $\hat{\gamma}_6$, which are of specific interest to us was closer to 10%). So our estimate of the overestimation of sampling variances, when using the bootstrap, is at least 10% for λ_1 with c =100. If the overestimation of variances is 10%, this corresponds to a sequence length of approximately 10/11 × 100 or $c \approx 91$. The $\log_{10}$ of 91 is approximately 1.959, and applying the approximation shown in figure 4.15, the probability of obtaining an incorrect tree has increased from 0.423 to 0.439 (or 0.404 to 0.420 if the calculations take into account covariances). This number is an underestimate of the bias in the bootstrap due to the non-linearity of the Hadamard conjugation, for this sequence length and in this example. However, as section 4.4.1 mentions, the bias in $\hat{\gamma}$ has practically disappeared by the time c =1,000, so we expect the bias in the bootstrap due to the extra non-linearity of the Hadamard conjugation, will also have decreased substantially.

The same sort of calculations can be used to predict how much improvement in tree selection will arise from using *rb* estimators (see section 4. 6.5, appendix 4.2). Using *rb* -42 (see figure A4.2.4) with c =100, should improve tree selection reliability to about that of a sequence of length 113 when using the logarithmic transform (this estimate is made assuming the variance of the relevant entries in $\hat{\gamma}$ will drop by 6%). This in turn should improve the accuracy of tree selection from 1 - 0.404 = 0.596 to about 1 - 0.382 = 0.619. Larger improvements will occur on trees with longer pathset lengths. It is important to check the accuracy of these estimates with simulations. The same issue of accuracy with short sequences, or about short internal edges, arises with most statistical tests in phylogenetics (e.g. significance of differences in likelihoods, Kishino and Hasegawa 1989, significance of the length of internal edges in a tree; with

likelihood, Felsenstein 1993, with distance trees where edge lengths are estimated with unweighted least squares, Li and Gouy 1991).

4.9 CONCLUSION

The aim of this study has been to survey the statistical properties of Hadamard conjugations as the foundation for putting this promising approach into a well understood statistical framework. We have been successful, for example, deriving the covariance matrix of $\hat{\gamma}$ and studying the form of its multivariate distribution. This allows the development of appropriate statistical tests of parameters such as edge lengths and their relative sizes for trees selected from $\hat{\gamma}$ (chapter 6). The observation that the distribution of $\hat{\gamma}$ seems to incur a low amount of bias even when relatively short sequences are analysed, adds confidence to the estimates it gives, and makes the construction of statistical tests on $\hat{\gamma}$ more straightforward. Other new results presented in appendix A4.2 suggest that an advantage will be gained by using the less biased estimator pathlength estimators of the same general type as pairwise distances derived by Tajima (1993a) (we expect the greatest advantage in cases where c is small and / or amounts of divergence are high). Things are now set to exploit this understanding of the statistical structure of $\hat{\gamma}$ in order to develop more rapidly converging and robust tree estimation procedures for sequence data transformed by a Hadamard conjugation (see chapter 5).

Another useful and pleasing finding is that the delta method introduces only a small bias when estimating the population covariance matrix from a random sample of sites. For Hadamard conjugations we recommend using the delta method in preference to the jackknife or the bootstrap when working with sampled data. In contrast, when doing simulations and knowing $\mathbf{s}(T)$ we would suggest caution when using $\mathbf{V}'[\hat{\gamma}]$ to estimate the population covariance matrix when either c is small and/or λ high. The ideal would be to compute the exact variance of ρ_i (noting that $\hat{r}_i$ varies as a function of a binomial distribution), yet this becomes computationally expensive as path sets sum more events in the sequences. In such cases we anticipate that approximating $\mathrm{Var}[\rho_i]$ with the variance of a logarithm-normal distribution will yield more accurate approximations than the delta method. That is: let the estimate of $\mathrm{Var}[\rho_i]$ be equal to the second moment of the distribution $\ln(N\sim(\mu, \sigma))$, where μ is r_i and σ the variance of r_i. One can be quite conservative in applying this improved approximation since the largest increase in accuracy, as well as the greatest saving computationally, will be in estimating the variance of the longest sets of paths. We would then calculate the covariances of ρ_i with other entries replacing the delta methods $1/r_i$ with the square root of $\mathrm{Var}[\rho_i] / \mathrm{Var}[r_i]$.

Our evaluation of the bias of the delta method for estimating the variance of genetic distance transformations appears to be unique in the literature. It is an important finding suggesting that

the delta method may be the method of choice (near minimum variance, and unbiased) for estimating variances of data after power transformations to remove the effect of multiple changes, given that biologists will almost always be working from a single sample. It will be interesting to see if the delta method turns out to uniformly give less biased estimates than the jackknife or the bootstrap when working with the various types of distance corrections discussed in chapter 3. The prediction would be that the bias of the bootstrap would be largest with the distances which make the largest difference between the observed and the estimated distance. These tend to be the more general distances such as the time reversible distance with a distribution of rates across sites, or the LogDet with removal of constant sites to compensate for rates across sites. For this same reason, we expect that using the standard bootstrap procedure of Felsenstein (1985) to estimate the support for internal edges of a tree is least flattering (i.e. most conservative) with deep internal edges in trees built from distances involving such transformations, and we have seen possible indications of this (e.g. the ultimately rapid drop in support for the archaebacteria monophyletic as invariant sites are removed, section 3.7.3).

The four state $\hat{\mathbf{s}}$ vectors contain 4^{t-1} entries, but the form of the Hadamard conjugation remains the same i.e. $\hat{\boldsymbol{\gamma}} = \mathbf{H}^{-1}(\ln(\mathbf{H}\hat{\mathbf{s}}))$. Therefore, the steps in computing the covariance matrix of $\hat{\boldsymbol{\gamma}}$ are the same, and the distributions of $\hat{\mathbf{s}}$, $\hat{\mathbf{r}}$, $\hat{\rho}$ and $\hat{\boldsymbol{\gamma}}$ are analogous to the 2-state case illustrated here. The major difference when analysing sequence data is that with 4-state data, the entries in $\hat{\mathbf{s}}$ become sparse with increasing numbers of taxa (t) much more quickly than in the 2-state case. As a result the vast majority of entries in $\hat{\boldsymbol{\gamma}}$ can have predominantly binomial forms (with expected values close to zero), an important consideration for statistical testing.

Simplifications of the full 4^{t-1} Hadamard conjugation, by 4^{t-1} or 2^{t-1} Kimura 2ST or Jukes-Cantor transformations, or the even simpler distance Hadamard, offer ways of both reducing the variance of marginal entries and making them more normal in shape. Each of these simpler transformations losses some of the information present in the full 4^{t-1} transformation, and it will be important to understand this information loss if these simpler transformations are to be used to help diagnose sequence evolution. In addition, the flexibility of the Hadamard conjugation framework shows itself, in that all these transformations can be made exact assuming a stationary distribution of rates across sites. They can all incorporate the constraints expected at the $\mathbf{r}$ or ρ level if a molecular clock holds, while the order 2^{t-1} 4-state Hadamard conjugations, and the distance Hadamard, also allow constraint to homogeneous Kimura 3ST and Kimura 2ST models. And of course all of these submodels, and various constraints, allow delta method estimates of the covariance matrix of $\hat{\boldsymbol{\gamma}}$.

There remains the problem that as the number of taxa increases, the number of entries in the covariance matrix of $\hat{\boldsymbol{\gamma}}$ which is $(2^{t-1})^2$ (or worse still $(4^{t-1})^2$ with 4 states) soon exceeds the storage capacity of all computers. We look forward to assessing methods (including approximations, e.g. see Szekely *et al.* 1993) that allow us to calculate just the variances (or covariances) of specific entries in $\hat{\boldsymbol{\gamma}}$, as an alternative approach around this problem.

The ease of estimating $\mathbf{V}[\hat{\gamma}]$, given $\mathbf{V}[\hat{s}]$, should also be of use to estimate the covariance matrix of $\hat{\gamma}$ when information about functional correlations of sites is available and can be used to directly modify $\mathbf{V}[\hat{s}]$. The main cause for concern is the inherent difficulty and, we suspect, also the high variance of estimating the correlation of sites which do not always covary (as even stem regions of rRNA allow some sites to have non-canonical pairing). Making such evaluations with real sequences should improve our understanding of how much correlation of sites may be affecting the variance of estimates of corrected distance values, and in turn the statistical efficiency of tree selection (as well as type 1 and 2 errors of hypothesis tests). This approach should also allow us to gain a better feeling for the types of systematic error that occur when certain sites have highly correlated evolution (Schöniger and von Haeseler 1994 have begun such analyses). Stuart and Ord (1990, p1076) offers useful insights as to how autocorrelation could be taken into account with the Hadamard conjugation, followed by tree selection.

Lastly, the analyses in section 4.5, showed the importance of considering the distribution of rates across sites to estimating a genetic distance. A related issue which was evaluated, showed that under simple models at least, large distances maintain a high degree of accuracy if there is no systematic error. The type of graphs used are suggested as a more suitable way to present information of sampling errors; their use certainly facilitates comparisons of expectations under different mechanisms of sequence evolution. The issue of data editing is seen to have great importance when there are some sites which have undergone very many changes along paths, e.g. as a neutral site would be expected to over the last 300 million years (when it may well have changed more than four times). Finding and removing these sites offers potentially great advantages to molecular phylogenetics of such old divergences, no matter what the method of analysis being used is (which includes maximum likelihood tree selection). From the point of view of systematic error also, these are the sites which are highly suspect.

Appendix 4.1 The calculation of HVH

Since **H**, **V** and $\mathbf{H}^t$ are all square (2^{t-1} x 2^{t-1}) matrices by ordinary matrix multiplication we require a total of 2^{t-1} x 2^{t-1} x 2^{t-1} = 2^{3t-3} operations for **HV** followed by the same again for this resultant square matrix × H giving a total of 2^{3t-2} operations. The fast Hadamard transform (a quicker way of multiplying by **H**, Hendy and Penny 1993, Tolimieri et al. 1989) may be applied to any row or column of a matrix multiplying with a suitable **H** matrix, and doing this will reduce the number of operations to $(t-1)2^{t-1}$ for that row, giving a total of $(t-1)2^{2t-2}$ operations for (**HV**) and $(t-1)2^{2t-1}$ for (**HV**)**H**. This is still computationally expensive, but possible for up to 11 taxa in less than a minute on a 486 computer with sufficient RAM to store a matrix the size of **V**.

Further due to the structure of **H**,

$$\mathbf{HVH} = \mathbf{vH}', \qquad \text{(A4.1-1)}$$

where **v** is **V** re-expressed as a column vector (the rows lined up one after the other, then transposed) and **H**' is the Kronecker product of **H** with itself, yielding a Hadamard matrix of the same type but with $(2^{t-1})^2 = 4^{t-1}$ rows.

The proof of 4.1-1 relies upon the fact that a 4^{t-1} Hadamard is the Kronecker product of a 2^{t-1} Hadamard multiplied by itself (Tolimieri et al. 1989). We partition **v** as **v** = [V*1, V*2, ... , V*N] where $N = 2^{t-1}$, i = 0, ... N-1, j = 0, ... N-1 thus $V_{ij} = \mathbf{v}_{i+Nj}$. It follows using this notation, that $\mathbf{H}_N\mathbf{V}\mathbf{H}_N = \mathbf{H}_{N2}\mathbf{v}$. The computational complexity of deriving $\mathbf{V}[\hat{\mathbf{r}}]$ via **vH**' remains unchanged from that in 4.1-1, with $(t-1)4^{t-1} = (t-1)2^{2t-2}$ operations when the fast Hadamard transform is used.

Calculating $\mathbf{V}[\hat{\rho}]$ from $\mathbf{V}[\hat{\mathbf{r}}]$ can be expressed as the multiplication $\mathbf{DV}[\hat{\mathbf{r}}]\mathbf{D}$, where **D** is a diagonal matrix with non-zero diagonal entries δ_i (for just this section A4.1) corresponding to the gradient at the point of transformation of r_i to ρ_i,. When working with $\mathbf{v}[\hat{\mathbf{r}}]$ this operation is efficiently performed by multiplying all rows then all column entries in **v** by δ.

So overall, $\mathbf{v}[\hat{\gamma}] = \mathbf{H}(\delta(\mathbf{Hv}[\hat{\mathbf{s}}]))$, where $\mathbf{v}[\hat{\mathbf{s}}]$ is a vector representation of the covariance matrix of $\hat{\mathbf{s}}$ as described above, and δ is the delta method applied to the appropriate entries in $(\mathbf{Hv}[\hat{\mathbf{s}}])$.

Appendix 4.2 An unbiased and reduced variance transformation of $\hat{\mathbf{r}} \rightarrow \hat{\rho}$

In this section we further evaluate the statistical properties of pathlength estimators, especially those that will estimate the pathlength in situations which would give a negative argument to the standard logarithmic correction. One of the problems encountered during simulations to study the sampling variance of $\hat{\gamma}$, was the occasional sample which gave a negative argument to the logarithmic transform of $\hat{\mathbf{r}} \rightarrow \hat{\rho}$. Since the logarithm of a negative number is not easily defined in such a situation, such samples were discarded. However when

studying real sequences with short sequences and large divergences, samples will be inapplicable with the standard logarithmic transforms. In this appendix we derive and study three new pathlength correction formulae. It turns out that in many realistic situations involving short sequences, one of them (the *rb* formulae based on the power series pairwise distance transformation formulae of Tajima 1993) has improved accuracy due to reduced sampling variance (a previously unknown property which apparently applies to this whole class of distance transformations).

A4.2.1 New pathlength transformations applicable when $\hat{r}_i$ is negative.

This section describes three new pathlength transformations which appear useful for inferring additive pathlengths when sampling error make the observed distance ≥ the maximum expected under the model. The first of these transformations, was inspired by Tajima (1993) who derives reduced bias (*rb*) path length correction formula for pairwise distances, applicable when the observed distance has a binomial marginal distribution. As he points out, these distance transformations remain valid for any observed distance. One of his distance transformations is based on the i.r. and i.i.d. 4-state Poisson model, and effectively replaces the logarithmic Jukes-Cantor transformation (to which it converges as $c \rightarrow \infty$). We may derive a distance transformation of this type to replace the logarithmic transform in both the 2-state and 4-state Hadamard transforms, by realising:

(1) Each entry $\hat{r}_i = 1 - 2\hat{s}_{J_i}$ is distributed as $1 - 2(B(c, s_{J_i})$ scaled by 1/c) (see section 4.2.3)(where $B(c, s_{J_i})$ denotes a binomial variable with parameters c and s_{J_i}, and s_{J_i} is an unweighted sum of specific entries in **s**). When working with a sample we will estimate $\hat{s}_{J_i}$ as $\hat{f}_{J_i}/c$ (or alternatively if we have already calculated $\hat{r}_i$, then $\hat{f}_{J_i} = c(1-\hat{r}_i)/2$).

(2) The path correction formula $-\ln(\hat{r}_i)$ may be reexpressed as $(1/b)-b(1-(\hat{s}_{J_i}/b))$ where $b = 0.5$, where b is the expected observed pathset length of random sequences under the 2-state Poisson model.

Thus, following Tajima (1993), our path correction formula $\hat{\rho}_i = \ln(\hat{r}_i)$ may be replaced with the reduced bias (*rb*) estimator,

$$\hat{\rho}_{(rb)i} = -\left(\frac{1}{b}\right)\sum_{m=1}^{\hat{f}_{J_i}} \frac{(\hat{f}_{J_i})^{(m)}}{mb^{m-1}c^{(m)}} = -\sum_{m=1}^{\hat{f}_{J_i}} \frac{(\hat{f}_{J_i})^{(m)}}{mb^{m}c^{(m)}}, \qquad \text{(A4.2.1-1)}$$

where $x^{(m)}$ (or $c^{(m)}$ or $\hat{f}_{J_i}^{(m)}$) is defined as $x(x-1)(x-2)...(x-m+1) = x!/(x-m)!$ In this equation c (the sequence length) is equivalent to Tajima's n, while $\hat{f}_{J_i}$ is Tajima's variable k, and m is equal to Tajima's i.

Here is an example of the application of formula A4.2.1-1, correcting an observed pathset length (let us assume it is a pathset between four taxa). In this example $\hat{r}_i = 0.2$, the sequence is of length $c = 10$ (so $\hat{f}_{J_i} = 10(1-(0.2))/2 = 4$), then following equation A4.2.1-1, $\hat{\rho}_{(rb)i}$ is the sum of ,

$$-\left(\frac{4}{1\times 0.5^{1}\times 10}+\frac{4\times 3}{2\times 0.25\times 10\times 9}+\frac{4\times 3\times 2}{3\times 0.125\times 10\times 9\times 8}+\frac{4\times 3\times 2\times 1}{4\times 0.0625\times 10\times 9\times 8\times 7}\right),$$

$$= -\frac{4}{5}-\frac{4}{15}-\frac{8}{90}-\frac{2}{105} \approx -1.17 \text{ substitutions per site.}$$

Since r_i is -2 × the expected length of the pathset, then the length of our quartet is estimated to be 0.585 substitutions per site (if we use the standard logarithmic transformation the length is estimated at -1/2ln(0.2) = 0.805). So overall, this gives,

$$\hat{\gamma} = \mathbf{H}^{-1}(rb(\mathbf{H}\hat{\mathbf{s}})), \qquad \text{(A4.2.2-2)}$$

(where *rb* is the reduced bias transformation of equation A4.2.2-1).

Next we introduce two other functions to estimate an additive pathset length, which are also applicable when $\hat{r}_i$ takes a negative value (in our model this is when *p*, the proportion of mismatches per site, takes a value of ≥ 0.5). These transformations are:

(A) M1: If *p* is ≥ 0.5 then reset *p* to p(M1) = (0.5/c + p_{max}), where p_{max} is the largest value that *p* can take from the sample at hand, such that $p_{max} < 0.5$. So if *c* was 20 and *p* was 12, then *p* is reset to p(M1) = (0.5/c + p_{max}) = (0.5/20 + 9/20) = 9.5/20 = 0.475.

(B) M2: If *p* is ≥ 0.5, then reset *p to* p(M2), such that the difference of p(M2) from 0.5 is $1/2^x$ times smaller than the largest p value which gives a defined logarithm of (1-2p). Here *x* is how many observed mismatch events must be deducted from p before the sample can be corrected by the standard logarithmic transform. That is, p(M2) will be,

$$\text{p(M2)} = \frac{\left(\frac{c-2}{2}\right)-2^{-(j-c/2)}}{c}, \; c \text{ even, or} \qquad \text{(A4.2.2-4)}$$

$$\text{p(M2)} = \frac{\left(\frac{c}{2}\right)-2^{-(j-(c-3/2))}}{c}, \; c \text{ odd,}$$

where *j* is the observed number of mismatches, and *c* is the sequence length. Following this rescaling of *p*, simply apply the standard natural log transform, so $\hat{\rho}_i$ = ln(1-2p(M*)). The methods M2 and *rb* both have the property that as the number of mismatches in the original sequence increases, so to does the inferred distance. Such a feature may be desirable (if it comes without unacceptably large errors) when the inferred distances are to be used to infer a tree. It is also worth noting that all these formulae are consistent under the 2-state Poisson model, in that as $c \to \infty$, these three pathset formulae all return the same value as the standard logarithmic formula (the consistency of which is proven in Hendy and Penny 1993).

A4.2.2 The contribution of bias to stochastic error in pathlength estimators.

This section studies the statistical properties of these three new formula in relation to the standard logarithmic transform, which is also the maximum likelihood pathset length estimator. Tajima (1993) cited bias as his main concern when deriving the reduced bias estimators. Earlier in section 3.3 we looked at how much bias was introduced into entries in $\hat{\gamma}$ due to the standard

logarithmic correction formula. This measured bias was put into perspective by considering its magnitude relative to the total sampling error. Our finding was that eliminating bias would result in only a small improvement in the accuracy with which $\hat{\gamma}$ could be inferred. What was clearly more important was to find a method which could reduce the overall variance. In passing it is useful to note that the ML estimator of pathlength can in fact be extended to infer the length of a pathset with an observed length of grater than 0.5. The ML estimate of the true pathset length when $\hat{p} \geq 0.5$ is on the boundary of the parameter space i.e. the distance is infinite. Unfortunately this is of no help, as it really gives us no useful information beyond indicating that the true distance is large (infinite distances are also liable to upset some tree reconstruction algorithms).

We begin by looking at the importance of bias in relation to overall sampling variance, for the four different formulae for inferring additive pathset lengths from observed numbers of mismatches. For these evaluations, it was possible to calculate all the required statistics exactly (i.e. without resorting to simulations), since the single variable $\hat{p}$ has a known binomial sampling distribution. All evaluations in this section are for the 2-state equifrequncy Poisson i.r. and i.i.d. model.

Table A4.1 shows the inferred pathset length for all possible observed pathset lengths with a sequence of length 20, and also gives the probability of each observed pathset length, if the expected pathset length is 0.39). As Tajima (1993) noted, the values estimated by the *rb* method can become very large (e.g. many thousands, as shown in table A4.2.1!). Notice that with such a short sequence, and large distance ($\delta = 0.75$), the binomial distribution of $\hat{p}$ results in a large probability (0.212) that a sample can not be transformed via the ML estimator to give a finite distance.

In comparison to the 4-state Poisson model, the 2-state model encounters bias, worsening signal to noise ratios, and inapplicable samples, at either shorter distances true, or shorter sequences (due to a higher probability of an observed substitution being reversed to a no observed substitution at a site due to further change). However, the 4 and the 2 state model are otherwise very similar and the results for the 4-state model translate directly to 4-state models. For example, the entries in the **r** vector of the 4-state Hadamard conjugation have the same binomial marginal distributions as the 2-state model and exactly the same transformation is applied. Further if we look at the Kimura 3ST formula (which can be though of as a Hadamard conjugation for 2 taxa), then we also see that the three terms in this formula, each involve taking the logarithm of a binomial variable. So indeed it turns out that if one is to study in detail the behaviour of these reduced bias transformations, it does not matter how many states one is looking at, but rather the relevant information is just the parameters of the marginal distribution of the binomial variable being transformed. So what we are revealing here is applicable to any reduced bias transformation of the form suggested by Tajima (1993), including the special cases he derived for use with 4-state sequences.

Table A4.2.1 Inferred values for our 4 pathlength estimators, for a sequence of length 20.
Inferred pathset length for all possible observed changes when the sequence length is fixed to 20. The last column gives the binomial probabilities seeing x changes when the expected number of observed changes is 0.39 (equating to a true distance of $\delta = 0.75$ under the Poisson 2-state model).

Obs. changes	*ln*	M1	M2	*rb*	B(20, 0.39)
0	0	0	0	0	0.0001
1	0.05	0.05	0.05	0.05	0.0007
2	0.11	0.11	0.11	0.11	0.0041
3	0.18	0.18	0.18	0.17	0.0156
4	0.26	0.26	0.26	0.24	0.0422
5	0.35	0.35	0.35	0.32	0.0858
6	0.46	0.46	0.46	0.41	0.1363
7	0.60	0.60	0.60	0.52	0.1731
8	0.80	0.80	0.80	0.66	0.1787
9	1.15	1.15	1.15	0.83	0.1513
10	na	1.50	1.50	1.07	0.1057
11	na	1.50	1.84	1.40	0.0611
12	na	1.50	2.19	1.92	0.0291
13	na	1.50	2.54	2.82	0.0114
14	na	1.50	2.88	4.64	0.0036
15	na	1.50	3.23	9.06	0.0009
16	na	1.50	3.58	22.51	0.0002
17	na	1.50	3.92	76.54	0.0000
18	na	1.50	4.27	383.08	0.0000
19	na	1.50	4.62	3142.44	0.0000
20	na	1.50	4.96	55571.19	0.0000

Table A4.2 shows the expected value, bias, and total sampling error about the true value, of the 4 transformations, for distances of zero up to 2.5 substitutions per site ($c = 20$). The column headed p_{inap}, is the proportion of random samples which will give an observed pathset length of ≥ 0.5, and are therefore cannot give a finite distance with the standard ML pathlength correction formula. Notice that with this very short sequence length, such samples are common if the true distance is 0.5 or greater. The next column shows the mean value of sample observed pathset lengths transformed by the ML method. This mean was calculated after excluding any observed distance which would not give a finite distance by the ML transformation (so it is the mean of just those samples which transformed to a finite distance). Notice how the ML transformation initially (for $\delta = 0.25$ and 0.5) gives on average an overestimate, but it then yields progressively larger underestimates of the true distance.

Next to the mean value of the ML estimator (table A4.2.1) we give the sampling error for the pathset length transformed by this measure. Here the sampling error is measured as the root mean square error, or RMSE. This is calculated as,

$$\text{RMSE} = \sqrt{\sum_{j=0}^{c} P(X = j)\left(\hat{\delta}_k - \delta\right)^2},$$

where j is, the observed number of mismatches, and P(X = j) denotes the binomial probability of j events, given c trials with mean cj. It is straightforward to separate the sampling error into two parts, since $\text{RMSE}(\delta) = \sqrt{(\text{Var} + \text{bias}^2)}$, where Var is $\text{E}\left(\hat{\delta} - \bar{\delta}\right)^2$ (the sampling variance of the estimator measured about the estimators sample mean), while bias is the squared difference of the sample mean and the population mean (μ) i.e. $\left(\bar{\delta} - \delta\right)^2$ (Stuart and Ord 1990, p.629, while Chapter 17, "Estimation", of Stuart and Ord, 1990, gives a useful discussion of desirable properties of statistical estimators). The number next to the RMSE, is the percentage of the of the RMSE that can be assigned to, or blamed-on, bias; this percentage is calculated as $1 - \text{RMSE}/\sqrt{(\text{Var})} \times 100\%$ (this is the amount by which the RMSE would decrease if all bias were removed).

Table A4.2.2 shows that for δ up to 0.75, removal of all bias from the ML estimator results in very little improvement in its accuracy. Consequently we say in this region of the sample space bias is not the problem (in spite of its serious sounding name, and some distinct differences between the true distance and the mean of the transformed distances).

Table A4.2.2 Expected values, bias and variances of path correction formulae for a sequence of length of 20.

δ	d_{obs}	p_{inap}	E[ML]	E[ln(m1)]	E[ln(m2)]	E[rb]
0.00	.00	*0.00*	0.00 (±0.00, 0.0%)	0.0 (±0.00, 0.0%)	0.0 (±0.00, 0.0%)	0.0 (±0.00)
0.25	.20	*0.00*	0.27 (±0.18, 1.0%)	0.28 (±0.19, 1.1%)	0.28 (±0.19, 1.1%)	0.25 (±0.15)
0.50	.32	*0.07*	0.52 (±0.27, 0.4%)	0.59 (±0.37, 3.0%)	0.60 (±0.41, 3.2%)	0.50 (±0.69)
0.75	.39	*0.21*	0.67 (±0.30, 3.6%)	0.85 (±0.44, 2.3%)	0.91 (±0.57, 3.8%)	0.75 (±4.66)
1.00	.43	*0.35*	0.75 (±0.38, 24.6%)	1.01 (±0.43, 0.0%)	1.13 (±0.64, 2.2%)	0.99 (±13.36)
1.25	.46	*0.44*	0.79 (±0.54, 47.3%)	1.10 (±0.44, 6.0%)	1.29 (±0.67, 0.1%)	1.22 (±24.11)
1.50	.48	*0.50*	0.81 (±0.74, 62.0%)	1.15 (±0.53, 19.8%)	1.38 (±0.70, 1.5%)	1.40 (±33.94)
1.75	.48	*0.53*	0.83 (±0.96, 71.1%)	1.19 (±0.68, 30.4%)	1.44 (±0.76, 7.9%)	1.54 (±41.55)
2.00	.49	*0.56*	0.84 (±1.20, 76.9%)	1.20 (±0.88, 36.4%)	1.48 (±0.87, 19.7%)	1.63 (±46.87)
2.25	.49	*0.57*	0.84 (±1.44, 80.8%)	1.21 (±1.10, 39.8%)	1.50 (±1.03, 31.5%)	1.69 (±50.40)
2.50	.50	*0.58*	0.84 (±1.68, 83.6%)	1.22 (±1.33, 41.9%)	1.51 (±1.21, 41.7%)	1.73 (±52.65)

Note: The symbol δ stands for the true pathset length, while d_{obs} is the observed pathset mismatch count (the Hamming distance for pairwise distances), and p_{inap} is the proportion of samples that will lead to invalid logarithm arguments. This is followed by the expected value of the logarithmic method, E[ML] (calculated after first neglecting samples giving inappropriate ln arguments). In brackets is the standard error of this method (RMSE), followed immediately by the percentage of RMSE due to bias. Then follow the same statistics for the other three transformations, M1, M2 and *rb*. For the last method (rb) the effect of bias upon the RMSE was always less then (0.005%) and for this reason is not shown.

Looking in turn at the other estimators represented in table A4.2.2, it is also clear that the contribution of bias to the size of errors made when using M1 to estimate distances of up to $\delta = 1.25$ is also minor. The main drawback of using M1 is the substantial rise in the variance of this estimator (relative to the ML estimator for distances of 1.00 or less). Notice also that with M2, bias is relegated to a very minor part of the total error for distances of up to $\delta = 1.75$. Unfortunately, again with this estimator the major problem is a much increased RMSE (relative to the previous two transformations), for δ up to 1.25.

The last columns in table A4.2.2 evaluate the *rb* estimator. Indeed, the *rb* estimator is nearly totally unbiased in this region of the sample space for distance of up to 1.00. However, the most interesting thing about this estimator is its initially low, then very large RMSE. As a consequence of the both the small size of the bias and the large size of the RMSE, the bias never appears to contribute to more than 0.005% of this estimators RMSE. The bias problem is solved, but at what a price! If these results are taken at face value, you would probably conclude that in this region of the sample space (and as shown later, in many other regions also) all three methods which aim to improve on the ML estimator by transforming very large distances, are failures according to their RMSE. This is because none of them has an average RMSE performance as good as that of the standard logarithmic transform, in the range of distances of real interest and application (the region where it is expected our models will be most accurate, given systematic errors in real data, will often be for $\delta < 1$).

A4.2.3 The reason for the often large RMSE of the *rb* estimator

How can the *rb* estimator have such a low bias but often very large RMSE? The source of the *rb* methods large RMSE (and also why M1 and M2 have larger variances than the log method for $\delta < 1$ in table A4.2.1) is that the method can return large, and more rarely very large distance values as table A4.2.1 shows. The occasional return of these very large distance values is also helps to reduce the bias of the *rb* method. Unfortunately, these large values reach a point where the square of their value increases more quickly than the inverse of the square of their probability of occurrence (which is a function of the tail probabilities of the binomial distribution), thus the RMSE goes up, and up and ... An alternative is to consider truncating some of these very large distances in an objective manner (perhaps up to the point where bias contributes no more than 5% of total RMSE). In practice, they will be truncated by biologists rejecting transformed distances of say 10 or more as showing saturation. Note also, that the standard methods of estimating variance which are based upon approximations (usually the delta method, Stuart and Ord 1987, see for example Kimura and Ohta 1972, Gojobori et al. 1990, section 4.2 of this thesis), do not take into account the rapidity with which some of these formula (especially the *rb* transformation), non-linearly increase in size. Later in section A4.2.6 it is shown that the delta method approximation for the variance of the *rb* method (Tajima 1993), often does poorly at estimating the RMSE of the *rb* method (and to a lesser extent the logarithmic transforms) when sequences are less than 100 sites long, and distances are in range of most interest (less than $\delta < 1$). For this reason, we recommend that any test of hypotheses of

transformed distances estimated with such short sequences be done with exact calculations (something easily accessible via spread sheets such as Excel)

On a more positive note there is also a hint of a promising feature of the *rb* estimator. Note that in table A4.2.2, at a true distance of 0.25 the *rb* method actually shows the smallest RMSE. In the next section we investigate both the effects of truncation and possible regions of reduced variance.

A4.2.4 The region where the *rb* estimator has the best RMSE

This section evaluates more extensively the accuracy of the transformations described above, again using the 2-state Poisson model, with all sites i.r. and i.i.d. Again sampling properties are calculated exactly, starting with the binomial distribution of the data. We are particularly interested to see how the *rb* method performs when it is truncated by rejection of the largest distances, and evaluate how much this reduces its RMSE. We define a set of *rb* estimators by how many of the most extreme observed distances are ignored; for example *rb*-8 is the same as the *rb* estimator, except all samples where the observed distance falls amongst the 8 most extreme possible values are discounted (e.g. with a sequence of length 20 *rb*-8 ignores any sample when more then 12 out of 20 sites are mismatches). We label these "conditional" estimators, because if they return a distance is conditional on that sample being less than some limit.

These "conditional" estimators involve a move back towards the problem that sometimes they will be inapplicable, leading to "throwing away samples". The probability of "sample rejection" is low for small true distances, but does become significant with larger distances. For example with a sequence length of 20, the probabilities of throwing away a sample when $\delta = 0.5$ is 0.002, when excluding the largest 8 observed distances, dropping to 0.0001 when excluding just the 6 most extreme distances, and so on (the eighth largest distance is transformed to 2.8 while an observed distance of 20 out of 20 returns $\delta = 55{,}500$!). If δ (the true distance) increases to 1 (keeping c at 20), then the probability of an observed distance being one of the 8 largest observed distances increases to 0.04 (or 0.004 for it being one of the 6 largest, and 0.0002 for being one of the 4 largest). It is hard to decide on an acceptable limit for the proportion of samples to be discarded, although one in a thousand would seem a conservative figure. At this level the -4 option is acceptable out to until about $\delta > 2$ (which is a large distance, given such short sequences). In the discussion below, it is indicated how many of the largest distances (s) can be ignored such that the total probability of rejecting a sample remains less than 0.001.

While RMSE is a common measure of accuracy, another measure which is often considered more robust because it does not increasingly weight large (but very rare) deviations from the mean is average absolute deviation (AAD). We will asses all the estimators by this measure as well, and expect it to be more favourable to the *rb* estimators. It is also worth noting that while ML estimators often have desirable properties such as the minimum possible RMSE errors, these properties are only expected to emerge asymptotically as $c \rightarrow \infty$. Here we show for the first time in phylogenetics, that consistent estimators of additive evolutionary distance can have smaller

errors than the ML logarithmic method as long as the true distance is less than a certain value (this cut-off distance being a function of sequence length). This cut-off value increases with sequence length and will rarely be limiting, as most of the useful phylogenetic information with short sequences will be contained in sequences diverged by less than this amount.

Figures A4.2.1, A4.2.2, and A4.2.3 show the results of our evaluations for distances of up to $\delta = 1.0$, with sequence lengths of 20, 40 and 100 respectively. The first part of this section considers in detail how different estimators performed for the shortest sequence length of 20, then the latter part looks at the trends that emerge with longer sequences. As figure A4.2.1 (which consider $c = 20$) clearly shows, the *rb* estimators (lines marked by squares) have markedly improved performance if they are truncated to ignore the largest observed distances (and in doing so incur a slight increase in the estimators bias, an increase more than offset by decreasing RMSE for realistic distances). Most importantly figure A4.2.1 shows that *rb* estimators have the sampling properties of all estimators (including the ML estimator) for δ small to moderate, by both the RMSE and AAD criteria. For δ in the region where the *rb* estimators perform better than the ML estimator ($\delta \approx 0.4$ when measured by RMSE, or 0.5 with AAD), then the probability of the observed distance being one of the six largest possible values is always less than 0.001. Thus, we say that the *rb*-6 estimator offers optimal performance for distances of up to 0.5, with little chance of a sample being rejected. The ML estimator tends to do better for this short sequence length, as δ becomes > 0.5. However, even in this region of above 0.5 it is not necessarily the best estimator since *rb* estimators ignoring just as many samples can have better performance (the relative performance of these two types of estimator for large (>1.0) distances can change, as the bias of these methods will then change with true distance, since they then effectively begin to take on a fixed mean value for all the larger observed distances). Its important to note that the AAD measure of sample error, while favouring the *rb* estimators slightly, does not contradict the trends evident with the RMSE measure. This suggests these are realistic measures of accuracy.

The M2 estimators initially perform like the logarithmic ML estimator, but then their sampling errors begin to get rather large (figure A4.2.1). Ignoring the very largest observed distances does not dramatically improve the performance of these estimators. This is because the source of most of their errors comes with the transformation of large, but not extremely large (say > 14/20) observed distances. Interestingly the RMSE and AAD curves for the M2 estimators get flatter with increasing distance, an effect which is more pronounced with the ML method, (the curve marked *ln(r)* in fig. A4.2.1). This factor is explored in more detail towards the end of this section. Figure A4.2.1(b) also shows the performance of the estimator M1. It does not perform particularly well, and unlike the M2 and *rb* estimators, does not have the appealing property of becoming larger as the observed distance increases.

(figure next)

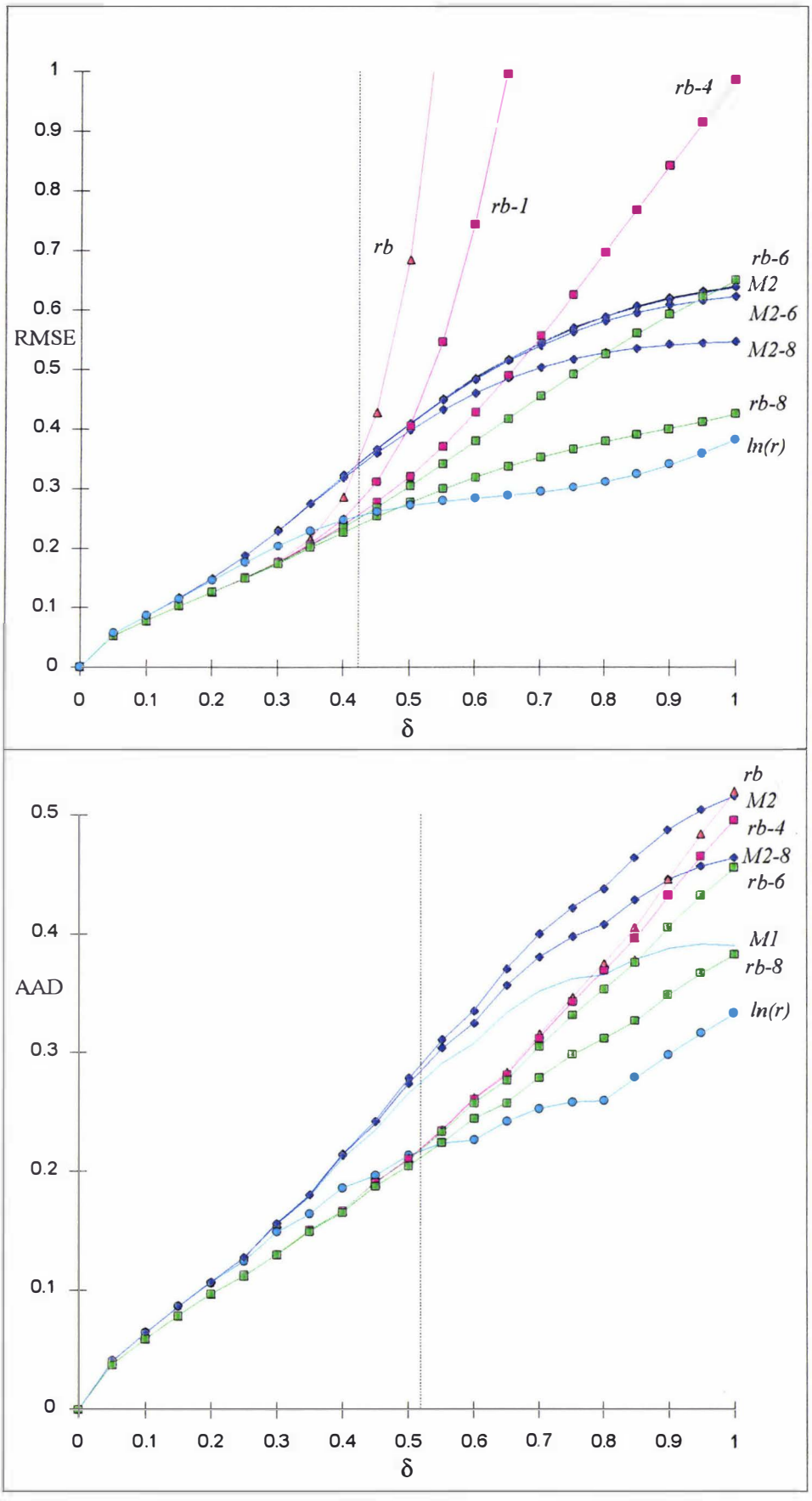


FIGURE **A4.2.1a** A plot of the true distance (δ), vs the RMSE of various distance transformations, with a sequence of length 20. A feature of special interest is that the lines with red, purple or green squares (truncated *rb* or reduced bias estimators) indicate better performance than the standard correction (the ML, or logarithmic estimator) indicated by the light blue circles. **A4.2.1b** A plot of δ vs the average

absolute deviation (AAD) of the different methods. The line with red triangles is the non-truncated *rb* estimator of Tajima (1993), while the blue diamonds indicate the transformation M2. The -x indicates how many of the most extreme observed distance values are being rejected with the truncated estimators. The vertical dotted lines indicate the distance up to which the moderately truncated *rb* method (-4) shows a clear advantage in sampling error over all the other methods. The truncation factor -4 is cited, as this is the truncation which within the range of values where it is better than the standard log transform, will not reject a sample more than 1 in 1000 trials.

One interesting effect shown in figure A4.2.1(b) is the bumpiness of the AAD, which is absent in the RMSE line. This effect is due to the small sample size, interacting with the discrete form of the binomially distributed observed distances. This effect diminishes as sequence length becomes longer (compare with figures A4.2.2(b) and A4.2.3(b)). This irregularity is not apparent when the RMSE measure is used, due to the smoothing effect of squaring, summing, then taking the square root.

We now examine the trends which emerge as longer sequences, specifically c = 40 and c = 100, are considered. As figures A4.2.2 and A4.2.3 show, the general performance of these transformations remains very similar. The "conditional" *rb* estimators have noticeably smaller RMSE and AAD than the standard ML transformation over a considerable range of δ (the area bounded by the line with solid circles and the lines with light grey squares). When c = 40, they are clearly the best estimators of d when $\delta < 0.6$ if measured by RMSE, or 0.72 in measured by AAD (fig. A4.2.2). For δ up to 0.7 the probability of having an inapplicable sample by the rb-16 method is always less than 0.0013, suggesting that this level of truncation is reasonable, and does not give rise to an excessive probability of not being able to transform a sample. As sequences get longer the *rb* estimators are the most efficient for increasingly wide range of distances, for example when c = 100, they are the best estimators up to approximately δ = 1.0 (see fig. A4.2.3). With c = 100, then with δ up to 1.0, the probability of having to discard a sample because d_{obs} was greater than 58/100 is less than 0.0011, confirming that the *rb*-42 estimator is indeed offering substantial improvement over the logarithmic method combined with a low probability of rejecting a sample.

(figure next)

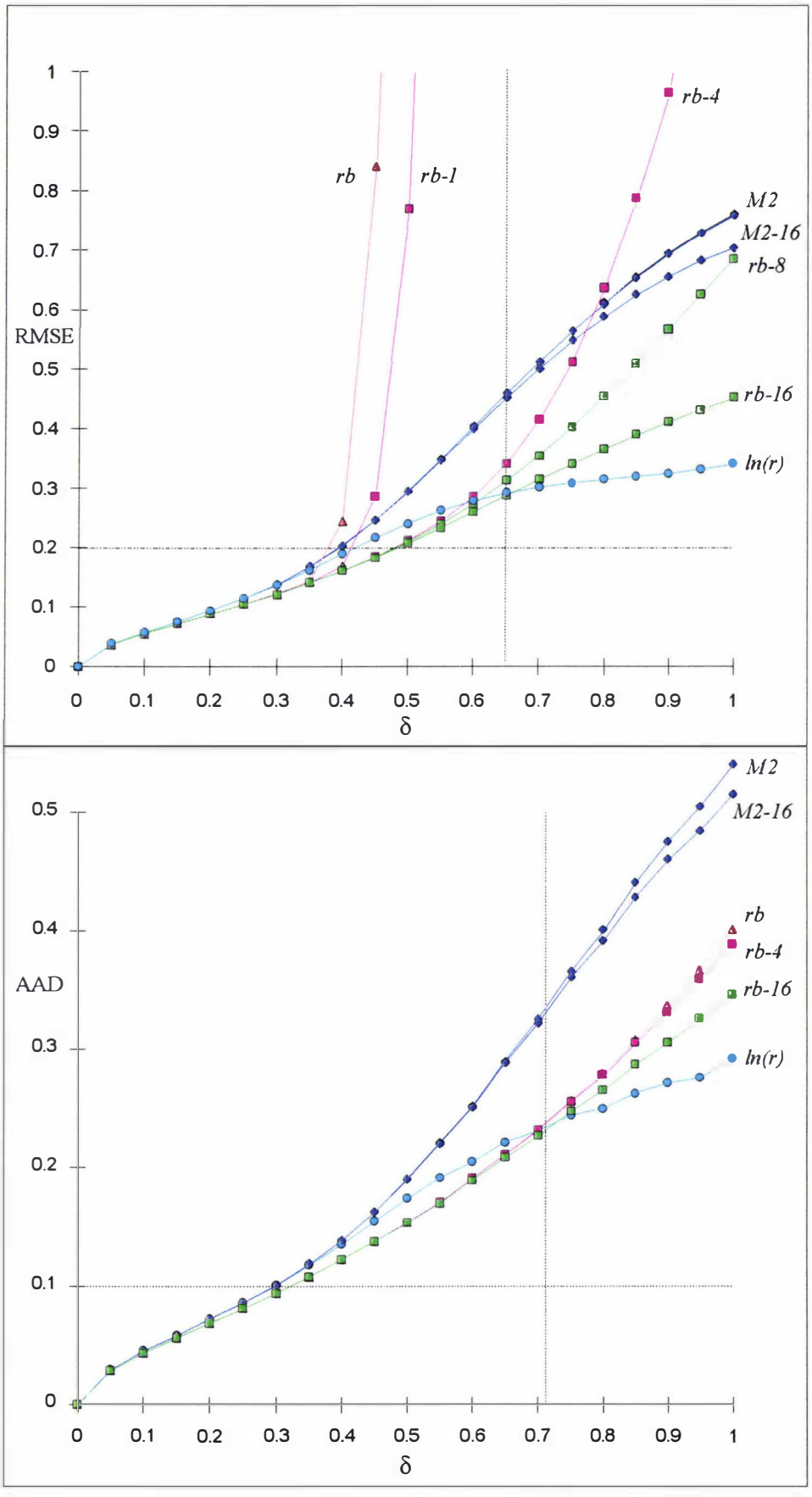


FIGURE **A4.2.2a** A plot of the true distance, δ, vs the RMSE of various distance transformations when the sequence length is 40 (transformation labels are as given in fig. A4.2.1). The region of particular interest is when the circles (the log transform) is higher than other transformations. **A4.2.2b** A plot of δ vs the average absolute deviation (AAD) for the different transformations. The vertical dotted line indicates the

value up to which the moderately truncated estimator (*rb-16*) performs better than the standard log transform. With a truncation factor of -16, this estimator has a maximum of a 1/1000 chance of rejecting a sample, for any value where it is better than the log transform.

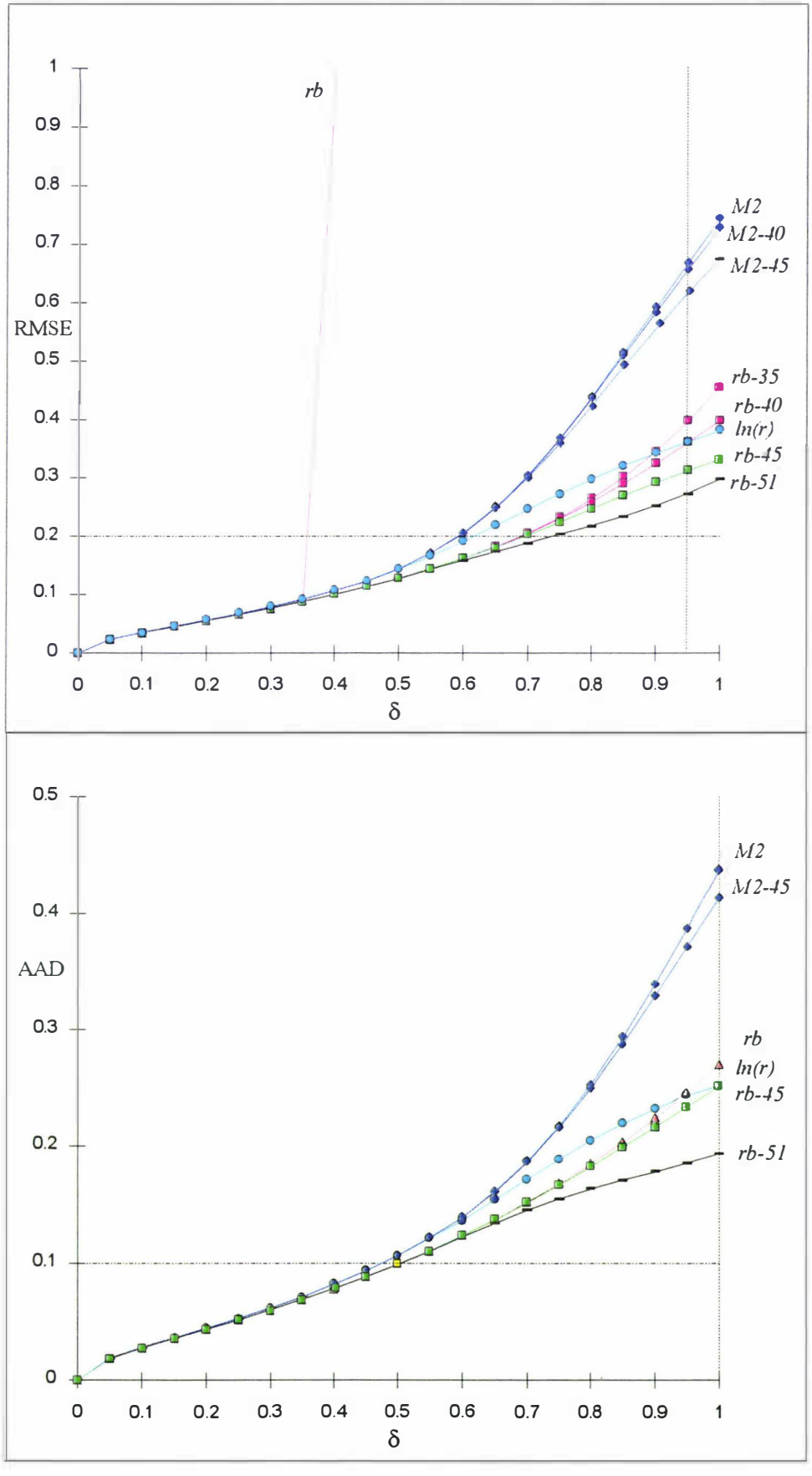


FIGURE **A4.2.3a** A plot of the true distance (δ), vs the RMSE of various distance transformations when the sequence length is 100 (transformation labels are as given in fig. A4.1). **A4.2.3b** A plot of vs the average absolute deviation (AAD) of the different transformations. The vertical dotted line is the value up to which a moderately truncated *rb* estimator (-45), has lower stochastic error than the ML ln transform, yet has no more than a 1/1000 chance of rejecting a sample.

It is informative to look at the relationship of the statistical efficiency of the *rb* estimators relative to the ML estimator with respect to δ and *c*. This trend can be clearly seen with the plot the proportional difference in the efficiency of these two estimators versus the true distance, δ. Such a plot is shown in figure A4.2.4 for the "conditional" *rb* estimators which all have a probability less than ≈ 0.001 of discarding sample for the range in which their performance is superior to that of the logarithmic ML method. Clearly these *rb* methods are not giving a constant amount of improvement over the ML estimator, but giving the greatest improvement at progressively larger values as the sequences become longer. Surprisingly, the maximum amount of improvement is also increasing with the sequence lengths studied. The amount by which the RMSE is lowered by replacing the logarithmic method with a "conditional" *rb* estimator can be compared to the amount the RMSE is expected to be lowered by increasing sequence length, *c*. (Here we use a large sample approximation to calculate the lowering of standard error with increasing *c*, which is also a reasonable guide in these small samples). A reduction of RMSE to 0.95 of its previous value is equivalent to increasing *c* by 11%, an improvement to 0.9 requires increasing *c* by 23%, and an improvement of 0.85 requires increasing *c* by 38%; all quite substantial amounts of extra information.

Converse to the advantages of *rb* estimators, the estimator M2's performance has clearly become worse as sequences have become longer. This deterioration of performance is especially prominent in the region where it is potentially of most use; at larger distances where samples of $d_{\bullet bs} > 0.5$ occur most frequently. It does not look to be a promising alternative to the logarithmic correction formulae for sequences of length 40 or longer.

Another point to notice is how the criteria AAD and RMSE agree more and more closely in their measure of the accuracy of these estimators as sequences get longer. By the time sequences are of length 100, only the "non-conditional" *rb* estimator still shows clear evidence of extreme distances affecting its RMSE measure. This trend is pleasing, in that it suggests that our conclusions are not dependent upon measuring the accuracy of an estimator with a particular statistic, which could artificially favour one method over another.

These results on the performance of *rb* estimators are very promising because it tends to be the inflating of overall variances when measuring larger values of δ which often result in poor tree section following transformation of the data, and not so much occasional exceedingly large distances (e.g. see section 3.7.4, chapter 5). Combined with the results in chapter 4, we anticipate that *rb* estimators will significantly improve the performance of tree selection on $\hat{\gamma}$ vectors from both distances and sequences, especially in the range of sequence lengths which biologists are frequently using (100-500 bp)(this expectation also goes for methods such as neighbor joining or

minimum sum of OLS branch length methods applied to transformed distances). The truncated *rb* methods should also see a significant improvement in the accuracy of edge weights calculated by these methods, an important consideration when wishing to infer divergence times (e.g. Waddell and Penny 1995). Additionally, *rb* estimators should reduce the small amount of bias that can appear in entries in $\hat{\gamma}$ (e.g. figure 4.4) to even more negligible levels.

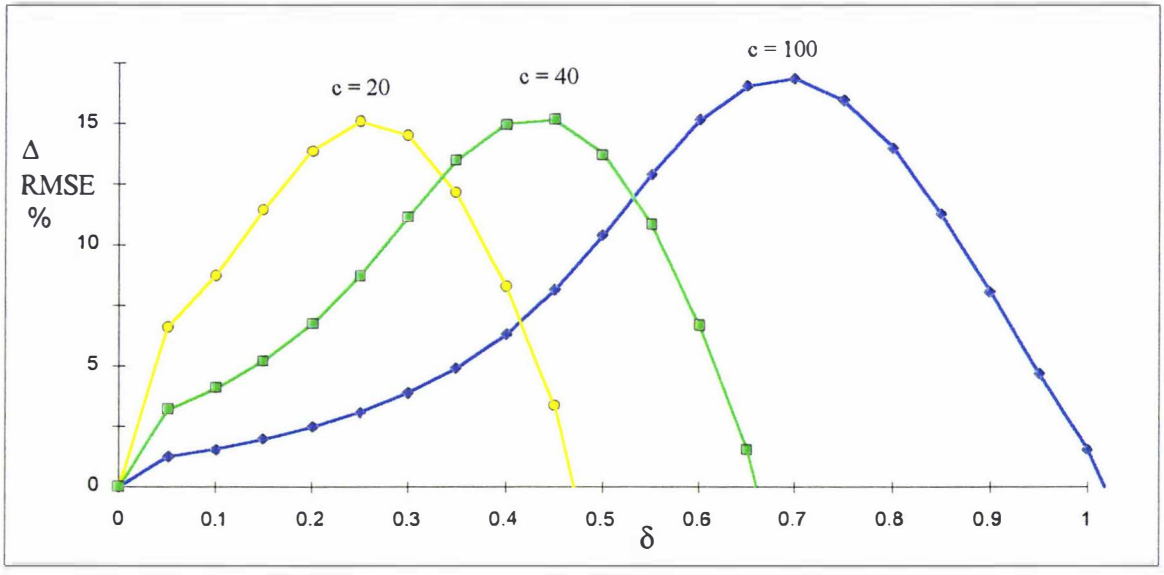


FIGURE A4.2.4 The amount by which truncated *rb* estimators improved upon the ML (logarithmic) estimator, plotted against the true distance, δ. The quantity Δ RMSE = (RMSE[ln] - RMSE[*rb*]) / RMSE[ln] × 100%, where RMSE[ln] is the RMSE of the logarithmic or ML method, and RMSE[*rb*] is the RMSE of a "conditional" reduced bias estimator. The "conditional" *rb* estimators shown are -8, -16, and -42 for sequence lengths of 20, 40 and 100 respectively.

A4.2.5 Accuracy of estimating large distances with small sequences.

Here we look briefly at how distance estimators perform in very small samples when the true distance is very large. While phylogeneticists should not be relying upon such data, such estimators may be of interest to geneticists trying to estimate how much change has occurred in a short section of DNA. Our interest is focused upon those estimators which rarely (if at all) throw out samples, specifically M1, M2 and the *rb* estimators.

The presentation of results in figure A4.2.5 shows that the M2 estimator performs surprisingly well for large distances between 1 and 1.8. Although the AAD is now large, surprisingly its ratio to the mean has dropped slightly. The AAD of this type of estimator has flattened out largely because the mean of this estimator which had been biased upwards, becomes close to the true value, before becoming an under estimate (see table A4.2.1) (which results in the AAD rising again). In figure A4.2.6, the performance of M2 estimators for short sequences and large distances using the RMSE criteria is shown. The results agree closely with those inferred form the AAD criteria, confirming that the result is not an artifact of the accuracy measure used.

Such large distances are unlikely to be of much use in phylogenetics because of their large relative errors, due both to sampling, and the high probability of large systematic errors. For

purposes of estimating quantities such as synonymous to non-synonymous rates, or absolute amounts of change, these crude estimates may give some clue. Clearly, however, even estimates for these purposes need to be considered carefully, from both the stochastic error and systematic error point of view. If no other information is available, an estimated distance of $\delta = 1.5$ made from just 20 sites (using the most reliable estimator in this region, M2), has a 95% CI from ≈ 0.26 to 2.54 (under the 2 state model)! With real data a large amount of systematic error could also cloud such an estimate.

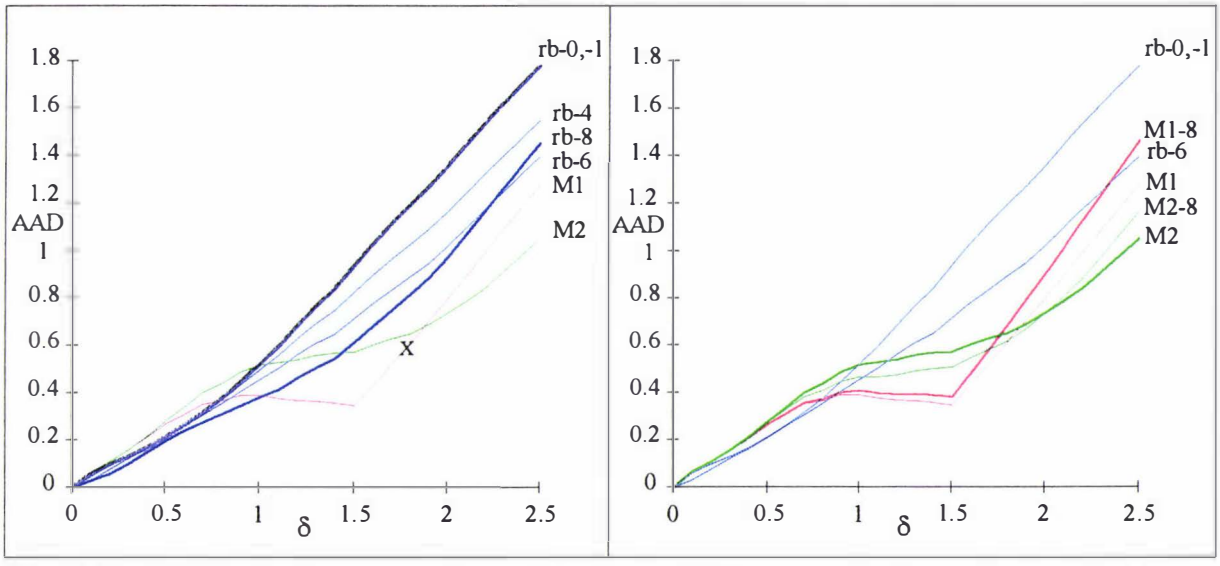


FIGURE **A4.2.5a + b** shows the accuracy of a wider selection of distance estimators for a short sequence (*c* = 20) (measured by AAD) for δ up to 2.5 (figure a shows more *rb* estimators, while figure b shows more logarithmic estimators). As seen earlier, when *c* = 20, most truncated *rb* estimators begin to perform worse than M2, for $\delta > 1.0$. If one wished to make distance estimates on small samples when the true distance was larger than 1.0, M2 shows the better accuracy of those estimators which do not throw out samples and continue to transform increasing large d_{obs} to increasingly large δ. However the overall impression has to be one of concern at the size of errors to be expected. As the sequence length increases, δ has to become larger before M2 transformations out perform *rb* transformations.

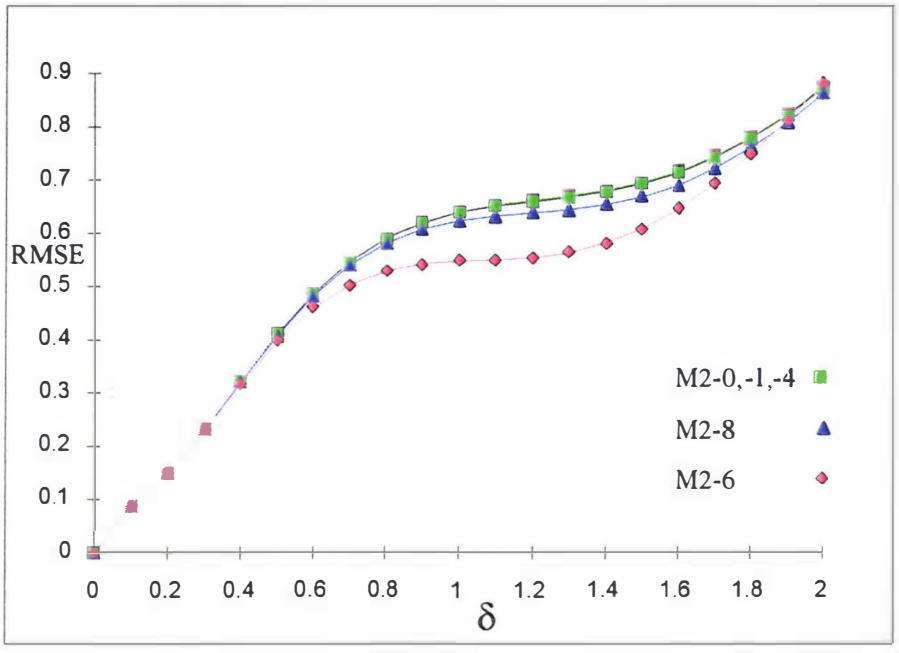


FIGURE A4.2.6 A closer look at the RMSE of the M2 estimators for a range δ up to 2 (c = 20). The RMSE of M2 estimators being near constant between δ = 1.2 and 1.5, is helped by their being close to unbiased in this region for this sequence length.

A4.2.6 Accuracy of delta method variance estimates for very short sequences.

Often in statistical tests of distances, assumptions of normality, and accuracy of the delta method approximation are assumed. Here the latter assumption is checked with exact calculations. In figure A.4.2.4 we have plotted delta method estimates of the variance of the Poisson estimate, and the reduced bias estimators (Tajima 1993) alongside the exactly calculated variances. For these evaluations the delta methods were applied to samples (not just the expected value of each distance), and this was done exactly by summing over all possible binomial sample distances. For short sequences (e.g. c = 20 or 40) the delta method approximations give both overestimates and underestimates of the true mean square error (figure A4.2.7).

(figure next)

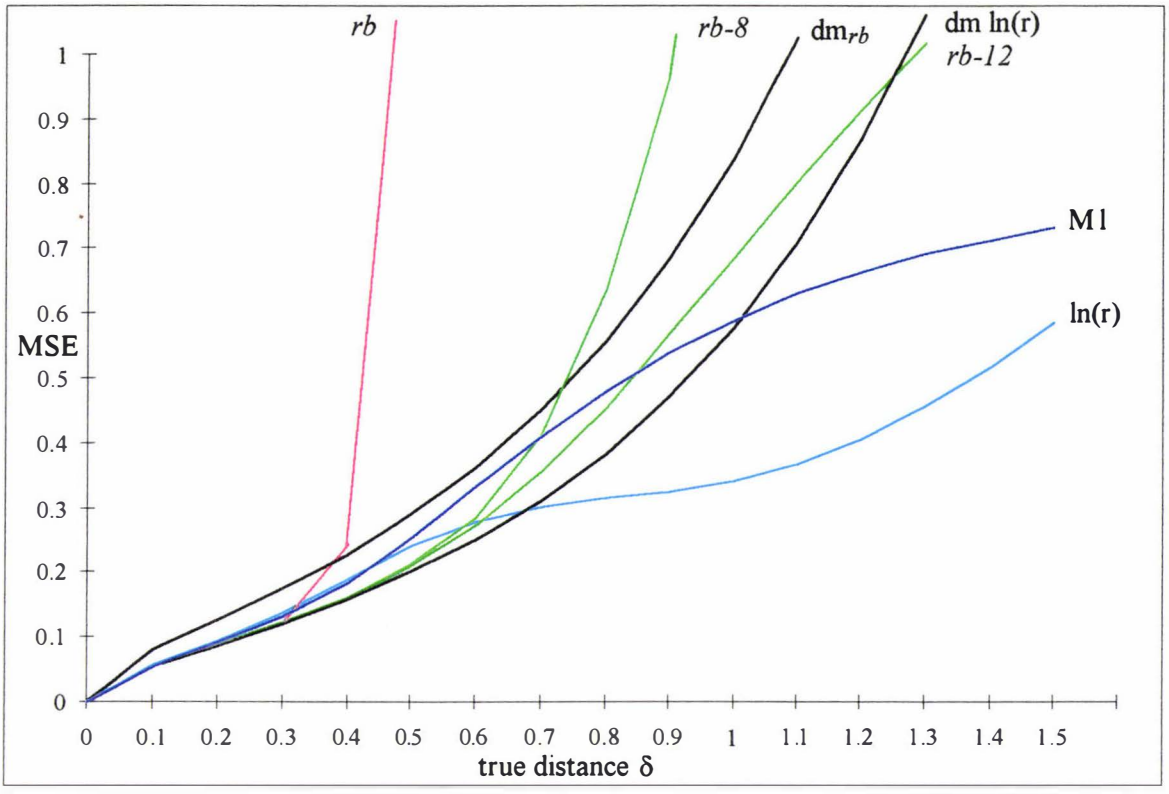


FIGURE A4.2.7 The mean square errors of various distance estimators compared to delta method estimates for c = 100. The green lines are reduced bias estimators with varying degrees of truncation, the blue lines are logarithmic corrections with either discarding all samples implying "infinite distance" (ln(r)), or else setting all such samples to the largest observed distance (M1), the black line is based on the standard delta method variance estimate of the logarithmic correction formulae, while the brown line is based on the delta method variance estimate given by Tajima (1993). Note, this is the delta method being used in practice upon samples, not just applied to the expected value of the observed distance, which in realistic situations is not available.

However as long as truncated *rb* estimators are bring used, and the true pathset length is not overly large (say less than 0.8), then the delta method gives are fairly consistent overestimate of the true MSE (averaging about 30% too large). For the logarithmic methods, the situation is reversed with the delta method underestimating the variance for distances up to about 0.7. Ironically the best combinations for overall accuracy in a certain range might be to use the logarithmic delta method variance formula to estimate the variance of *rb* estimators up to a distance of about 0.7, but use the *rb* variance estimate for M1 distances up to about 0.7. For c = 100, the delta method variances applied to samples perform markedly better as indicated in earlier simulations in chapter 4.

A4.2.7 Discussion

Here we see that the type of estimator derived by Tajima (1993) is useful, but for reasons different to those suggested by Tajima. From the results here *rb* estimators look a promising way of improving the accuracy of genetic distances, inferred from small sample sizes. Further more following on from results (1) and (2) in the first paragraph of this appendix, Dr Mike Steel has derived reduced bias estimators for Hadamard conjugations when there is a continuous distribution of rates across sites (e.g. the Γ distribution). Thus the benefits of reduced bias estimators need not

be associated with only the most simple models of path length and distance correction, but may extend to more general models, and also distance estimates. Their application to the LogDet or the generalised time reversible distance, requires careful consideration, since the distribution of the quantities to which the log function is applied (the determinant, and individual eigen values of **P**, respectively) does not have a simple binomial, or multinomial distribution.

Another avenue we are presently exploring is "what is the accuracy of truncated series *rb* estimators?" It turns out that when an rb estimator transforms a overly large observed distance, into a very large transformed distance (e.g. $\delta = 5$ or more), it is the last terms in the series (A4.2.2-1) that contribute most to such extreme results. There may be some formula to truncate this series, so as to yield much improved accuracy, and perhaps without needing to condition on rejecting even 0.1 % of samples. We are presently exploring such alternatives. Another area which we are also actively exploring is just how much improvement in tree selection the *rb* estimators can offer over the standard ML estimator (and how they compare with using just the observed distance, p). We can make a crude estimate of how much *rb* estimators reduce the sampling variance of $\hat{\gamma}$ by using a refined form of the delta method to predict $\mathbf{V}[\hat{\gamma}]$ the when they are used. Firstly, estimate the variance of entries in $\hat{\rho}$ by exact calculations (given the observed distance from $\mathbf{r}(T)$), then estimate the covariance of $\hat{\rho}_i$ and $\hat{\rho}_j$ as,

$$\text{cov}(\hat{\rho}_i, \hat{\rho}_j) = \text{cov}(\hat{r}_i, \hat{r}_j) / [(\text{s.d. } \hat{\rho}_i / \text{s.d. } \hat{r}_i) \times (\text{s.d. } \hat{\rho}_j / \text{s.d. } \hat{r}_j)]. \qquad \text{(A4.2.5-1)}$$

Doing these calculation for all pairs of transformed pathset lengths an estimate $\mathbf{V}'[\hat{\rho}]$ is obtained, which then need only be transformed by equation 4.2.5-1 to give $\mathbf{V}'[\hat{\gamma}]$.

There is a further piece of information that may be useful for refining methods to infer an additive distance, when the observed distance is greater than the largest observed distance expected under the model. If it is assumed the model of sequence evolution followed some known or unknown tree, then an upper bound on the largest path through a tree is that it can be no larger than twice the second largest distance through the tree. Thus, if just one distance out of a matrix of n(n-1)/2 distances gives an invalid argument for a power transformation, logically it should not be set to a value any larger than twice the next largest value in the matrix of transformed distances.

To conclude, the property of a substantial reduction in RMSE for *rb* estimators is very appealing. Mike Steel (pers. comm.) reports that it is straightforward to modify any of these estimators to allow for a distribution of rates across sites. Doing this for, say a Kimura 3ST model, and combining such a distance with the "infinite distance" modifications of section 3.3.2 (e.g. equation 3.3.2-2), could yield a very useful distance. It's properties for reduced sampling variance, and improved robustness would deserve careful consideration against other distance estimators that perform well when base composition is non-stationary (especially against the invariant sites-LogDet transform). Such distances would also find a natural place for use with the order 2^{t-1} 4-state Hadamard conjugations of appendix 2.6. As with standard logarithmic distances, it will still be important to consider weighting transversions over transitions in some

situations, when it can be shown the transition component of the distance has a higher variance than that of the transversion component.

CHAPTER 5:

PROPERTIES OF TREE SELECTION CRITERIA

5.1 INTRODUCTION

In this chapter we will consider the statistical properties of criteria that may be used in selecting a tree. A major focus is on tree selection from the vector $\hat{\gamma}$, but we frequently compare it with selection from $\hat{\mathbf{s}}$ (observed sequences), and from both transformed and untransformed distances. Wherever possible use will be made of statistical theory, contrasting the known properties of tree estimators with the statistical properties of $\hat{\gamma}$, as described in chapter 4. Properties considered include rate of convergence to the true tree, bias (with respect to recovering the true weighted tree) and consistency (with respect to recovering the correct unweighted tree). Systematic errors in edge length estimation will be of particular interest to biologists making inferences from the weighted tree, particularly features such as relative divergence dates, and relative rates of evolution. Consistency in estimating the unweighted tree is of interest to all biologists. When the underlying assumptions of the mechanism of substitution are progressively violated, the ability of a tree selection criterion to minimise edge length bias, and remain consistent in recovering the unweighted tree may be considered measures of its robustness. Also considered is the hypothesis that systematic errors in estimating edge length (a form of bias) eventually leads to inconsistency of unweighted tree selection in some part of the parameter space. Each tree selection criterion is illustrated with an application to a real data set.

A question of contemporary interest, is the nature of the relationship between tree selection from $\hat{\gamma}$ and maximum likelihood (ML). Hendy (1989) has suggested, for example, that selecting a tree from $\hat{\gamma}$ using the closest tree criterion is somehow equivalent to using the ML tree selection criterion of Felsenstein (1981a), which is applied to the vector $\hat{\mathbf{s}}$. Here we explore this issue in more detail.

It is straightforward to calculate the likelihood of a given set of data by using a Hadamard conjugation to infer the probabilities of different sequence patterns for a specific weighted tree (note that likelihood of a site pattern is just its probability under an i.i.d. model). Using Hadamard conjugations, such calculations can be made surprisingly quickly for up to ≈ 20 taxa, irrespective of whether a continuous distribution of rates across sites is modeled. Methods of maximum likelihood with a distribution of rates across sites have been independently developed by the author (see Steel *et al.* 1993c, Waddell and Penny 1995, and Yang (1993, 1994b). A second area of topical interest is whether there is only a single ML optima per tree and just as importantly how can it (they) best be found. Towards this purpose we derive the Hessian matrix (directly related to the Fisher information matrix, Stuart and Ord 1990) of the likelihood function with respect to the edges in a tree, and consider other related conjectures.

In the second half of this chapter we use exact (asymptotic $c \rightarrow \infty$) numerical calculations to explore the robustness of different tree selection methods (mixtures of selection criteria and mechanisms for correcting for multiple changes) when the assumption of identical rates is violated. This allows us to test some theoretical predictions from the first half of the chapter. We compare most of the commonly used tree selection criteria including parsimony, maximum likelihood and fit to distance matrices, as well as criteria for selecting trees from $\hat{\gamma}$. It is shown that all tree selection criteria can become inconsistent when rates across sites are unequal, and that this can also occurs under a molecular clock (an example from these studies appears in Lockhart *et al.* 1995).

It was shown by Felsenstein (1978a) that parsimony and compatibility applied to the observed sequence patterns, can become inconsistent estimators when rates of change are unequal in different lineages. Hendy and Penny (1989) generalised Felsenstein's observations, and showed that even under equal rates per unit time (a molecular clock) parsimony and compatibility are inconsistent when applied to the observed sequence patterns. They described the parts of the parameter space where these two methods become inconsistent as "long edges attract" (Hendy and Penny 1989). Others have shown that distance based tree selection criteria can also be inconsistent if rates of change across sites are unequal (Olsen 1987, Jin and Nei 1990), while chapter 2 showed similar problems could occur with data transformed by the Hadamard conjugation. Here we show that a "long edges attract" problem can also occur with maximum likelihood methods. From this it is apparent that a reformulation of "long edges attract", identifies it as an Achilles' heel of all presently used phylogenetic methods.

It is also important to consider what overcorrecting the data can do. Here we show that there is an "anti-Felsenstein zone" and we evaluate a wide range of tree selection methods in it. The anti-Felsenstein zone is a particular case of a "long edges repel" problem (section 3.4.3), and alternative names for it could be the "inverse-Felsenstein zone" or the "opposite-Felsenstein zone". This zone posses a serious problem for real analyses, especially if the model / data relationship is not understood well. The "anti-Felsenstein zone" problem has important consequences also for simulation studies such as those performed by Huelsenbeck and Hillis (1993), which may need re-interpretation. Methods of analysis must be considered to have two potential Achilles' heals, neither of which can be excluded without study of the true processes of evolution of the data. We also show that different trees can give the same sequences, so that without knowing the model, you could not identify the true tree. This possibility was independently identified by Steel *et al.* (1995) who predicted such problems with an existence theorem.

A series of evaluations and analyses of maximum likelihood with unequal rates across sites is given. Included are the studies of hominoid sequences reported in Waddell and Penny (1995), and also studies of some ancient rRNA. The latter case appears to be a good example of where ML, allowing for unequal rates across sites, strongly supports one tree (the eocyte tree), while a goodness-of-fit test suggests the model is reasonable, yet the tree is probably incorrect (the correct tree identified in chapter 3 by the slowest evolving sites is the archaebacteria). This may

be an example of an anti-Felsenstein zone problem with real data. Clearly there is much need to consider the true processes of evolution when studying ancient divergences.

Also considered are extensions of ML models to model situations where sites do change their relative rates (a form of covarion model, related to those of Fitch and Markowitz 1970), and to situations where sites do not have the same history due to reticulate phylogenies. A particular example of "reticulate phylogenies" is the case of ancestral polymorphism giving trees which are different to the species tree. We develop an ML model of this and applying it to primate sequences obtain an estimate of the genetic diversity of our ancestors just prior to the divergence of human and chimp lineages. This estimate of ancestral diversity then allows an estimate of the population size of these proto-Australopithecines. This turns out to be a credible 100,000 to 300,000, which is quite similar to that of modern African ape subspecies.

5.2 TREE SELECTION OPTIMALITY CRITERIA FOR $\hat{\gamma}$

In this section we describe an array of optimality criteria that may be applied to transformed data, in order to select a tree. Many of them have well understood similarities to traditional statistical estimators, and the insights that these offer in to the properties of these methods are discussed. However because they are being applied to a tree, some of there properties do not necessarily match up to known descriptions of statistical estimators. By analysing a real data set we see some of these "tree" properties emerge more clearly. One particularly interesting insight is that ML estimation on $\hat{\gamma}$ can be equivalent to ML estimation on $\hat{s}$, and this allows us to discuss the relationship between the ML procedure of Felsenstein and various ways we may select a tree from transformed data (both $\hat{\gamma}$ and pairwise distances).

5.2.1 Some important properties of $\hat{\gamma}$ with respect to tree selection

Before proceeding it is worth reiterating some properties of $\hat{\gamma}$. Firstly, γ_0 is the negative sum of all other entries in γ, and as such is not an independent quantity. Indeed all its statistical properties are just the "sum" of the properties of the independent variables. In addition, it cannot be selected as an edge in a tree because it does not correspond to any particular edge. For this reason it is "neglected" when searching for a tree (by neglected we mean it may be of use, but only as the sum of other variables, not in its own right). Its existence is guaranteed because it is a by-product of any vector multiplied by **H**. Its statistical non-independence arises from the fact that a probability vector like **s** must sum to one, and thus has one less free parameter than its number of entries.

An important issue which quickly arises is how to interpret negative values in $\hat{\gamma}$. It is certain that the true tree has all edges positive, and thus $\gamma(T)$ is a vector where γ_i is always positive (excluding γ_0). We say that $\gamma(T)$ lies in the positive quadrant. If a negative entry is encountered in

$\hat{\gamma}$ then its relationship to tree selection needs to be dealt with. The most direct implication of a negative entry is a lack of evidence for that entry being greater than zero in $\gamma(T)$ under the model. Such a lack of support may occur because of a lack of substitutions, due to the data coming from a different model (see chapter 2), or because of sampling error (see chapter four). Since $\gamma(T)$ cannot take on a negative value, it is often considered that the cost of such an entry can be effectively set to zero and this is the minimum edge weight we will accept in a tree. Hendy and Penny (1993) effectively took this approach as shown later.

As chapter 4 suggested, if negative entries in $\hat{\gamma}$ are considered to be solely due to sampling error, then there are situations where our best guess as to the identity of an internal edge is that entry with the least large negative value. It is not the same criteria as saying "accept a negative edge if it gives the best overall fit" (e.g. Farris 1981, 1985, 1986), but rather says "if there is no choice because all resolutions of an internal edge are negative, choose that which is least negative, i.e. the edge which is closest to positive." (Felsenstein 1984, 1986 considers other ways of resolving this dilemma). This criteria can be applied to other tree estimation procedures, not just trees selected from $\hat{\gamma}$.

If we use a method which assumes independence between all entries in $\hat{\gamma}$, then edges with a negative weight should not influence tree selection if they are not compatible with the optimal tree. Accordingly, it is usual to set them to a nominal value of zero and only reconsider them if edges in the tree cannot be resolved with a positive edge weight. Only occasionally may they need to be reconsidered in resolving an internal edge and given that a negative entry must be chosen, then it would make sense to choose that entry which was the fewest standard deviations from zero. The edge weight in the tree is then most sensibly set to the ε (a tiny positive number). With methods such as generalised least squares tree selection (GLS) which takes into account correlations between edges, then negative entries are always considered. They will influence the tree selected in complex ways which will vary depending upon the covariance matrix, the entries in $\hat{\gamma}$, and the maximum number of edges in the tree.

Lastly, it is useful to remember that $\hat{\gamma}$ is a sufficient statistic of the vector $\hat{s}$ (the rescaled observed data). By sufficient we mean that it contains all the information in $\hat{s}$, which is easily shown by the fact that the inverse Hadamard conjugation recovers the original data exactly (as long as all $0 < r_i \leq 1$). This throws light on the important question of the relationship between tree selection on $\hat{s}$ versus tree selection from $\hat{\gamma}$. Combined with the fact that the maximum likelihood point is invariant upon transformation (Stuart and Ord 1990, chapter 18), then we can say that the maximum likelihood point in $\hat{s}$ (under, say, a multinomial model) will be the same as that in $\hat{\gamma}$. The maximum likelihood point in $\hat{\gamma}$ can be searched for directly if there is an accurate description of its sampling distribution. The main advantage of this finding is that it allows us to make some pertinent observations about why ML on $\hat{s}$ may do so well compared to tree selection criteria applied to transformed data.

5.2.2 Some real data to illustrate tree selection criteria

For this section we will use similar data to that in figure 2.7. The four sequences are the transversional changes in 16S-like rRNA from the same taxa, except that the sequence of *Halobacterium volcanii* (now sometimes assigned to the genera *Haloferax*) has been substituted with that of its relative *Halobacterium salinarium*. There is a further difference in these data sets because they come from different alignments; this latest data set is from the alignment of Dams *et al.* (1988), and is also used by Churchill *et al.* (1992) in their paper on sample size effects. In order to evaluate some of the differences in these data sets, the extended Hadamard conjugations, allowing for different distributions of rates across sites, were applied to each. Figure 5.1 shows the results which may be compared to those of the data aligned by Lake (1987) and shown in figure 2.7. The data set from Dams *et al.* is useful because, (1) it allows us to use the Hadamard to visualize differences in signal due to what Olsen and Woese (1989) describe as an alignment bias in Lake's data; (2) the data does not appear to fit the model as well as with the Lake alignment (see figure 2.7) which better highlights some differences in tree selection criteria; (3) it allows us to compare some of our results with those obtained by Churchill *et al.*(1992).

Figure 5.1 indicates that under this alignment, the data do not fit the model quite as well as with the alignment of Lake (1987). This data still shows better support for the "eocyte tree" when the model fits best (i.e. both γ_3 and γ_5 near zero). An obvious question is: does this mean that the "eocyte" tree is favoured by this trees from 16S-like rRNA? Not necessarily. Firstly, as we show later, the statistical support for the eocyte tree over the archaebacterial tree is near zero (i.e neither tree is significantly better supported than the star tree). Secondly, a result in Olsen and Woese (1989) appears to be relevant. Specifically they found that Lake's (1987) method of evolutionary parsimony tends to favour the eocyte tree for most reasonable alignments if the sequence from the methanogenic archaebacteria is a halophile (but otherwise not). Evolutionary parsimony tends to place heavy emphasis upon transversional changes, so our result may be a reflection of the same underlying feature of the data. It tends to illustrate one of the dangers of using just four species at a time to infer a phylogeny, since there are not additional taxa to break up long edges, or hopefully contradict strong biases. Note that this support for an eocyte tree is not generally apparent in the 28 taxa 16S-like rRNA data set analysed in chapter 3, probably largely due to the inclusion of other non-halophilic methanogenic bacteria.

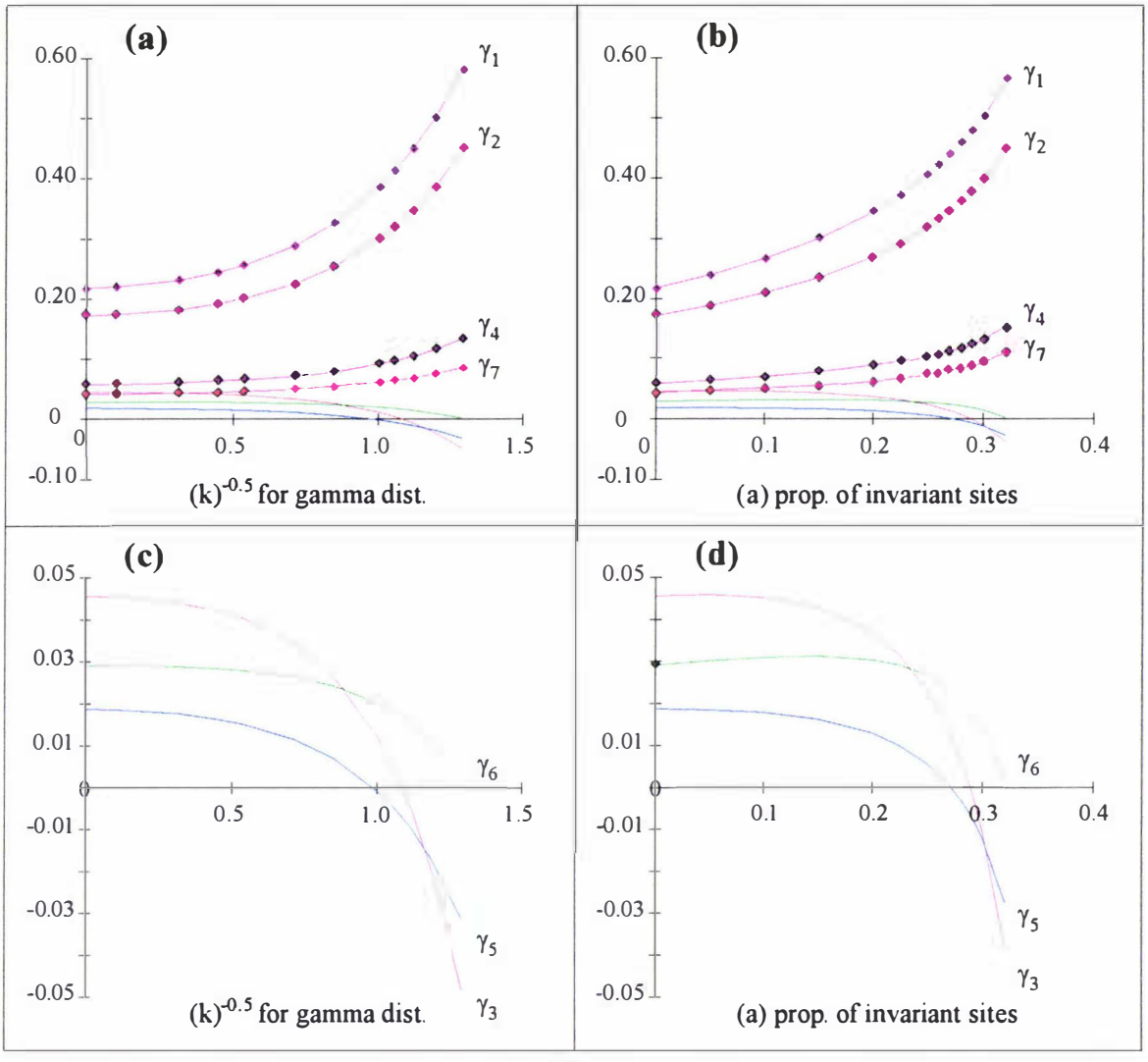


FIGURE **5.1a-d** Correcting transversional changes in ancient 16S-like rRNA for multiple substitutions. Values of entries in $\hat{\gamma}$ (γ_0 excluded) after correction is made assuming different distributions of rates across sites. The data are the transversion patterns in the 16S-like rRNA of four taxa from Dams *et al's* (1988) alignment as used by Churchill *et al.* (1992). The taxa are: (1 = γ_1) the eukaryote, *H. sapiens*; (2,γ_2) the eubacterium, *E. coli*; (3,γ_4) the halobacterium, *Halobacterium salinarium*; (4, γ_7) the sulphur metabolising bacterium, *Sulfolobus solfactarius*. **5.1a + c** Corrected sequence patterns assuming a gamma (Γ) distribution of rates across sites plotted against the coefficient of variation of this distribution (= the inverse of the square root of the shape parameter k). **5.1a** Entries in $\hat{\gamma}$, with just those pertaining to external edges labeled, while in **5.1c** just $\hat{\gamma}$ relating to internal edges on a tree. **5.1b + d** as for (a and c) but the proportion of invariant sites modeled with parameter 'a'. In both cases the apparent support for an archaebacterial tree ((1) and (2) together, γ_3) begins the highest, but then falls away steadily as the variation of rates across sites increases, eventually leaving the eocyte tree (γ_6) as favourite until all binary tree's lose support. The $\hat{s}$ vector for this data is [787, 178, 145, 68, 56, 33, 40, 45]. Comparing this figure with 2.7 (in chapter 2) notice how plotting the x-axis as coefficient of variation of the underlying gamma distribution (rather than 1/k) makes the plots for the data transformed assuming a gamma distribution look more like those obtained by removing a proportion of unvaried sites.

5.2.3 Ordinary (or unweighted) Least Squares (OLS).

Ordinary or unweighted least squares criteria will be familiar to those who have done linear regressions. OLS is a maximum likelihood method when all data points are multivariate normally distributed, have equal variance, and have all correlations equal to zero. As chapter 4 shows, all of these assumptions can be violated to varying degrees in $\hat{\gamma}$ (and also $\hat{s}$, $\hat{r}$, and $\hat{\rho}$). Traditionally statisticians regard the OLS criterion as robust to small violations of the assumptions, however tree selection is not such a traditional approach and violations to the model can be considerable. Our search method seeks to minimise the sum of squares (ss) of γ_i entries excluded from the tree,

$$\begin{aligned} \text{OLS SS} &= \sum_{i \in T} \left(\hat{\gamma}_i - \gamma_i(T)\right)^2 \qquad (5.2.3\text{-}1) \\ &= (\hat{\gamma} - \gamma(T))^T (\hat{\gamma} - \gamma(T)), \\ &= \sum_{i \in T} \hat{\gamma}_i^2 \end{aligned}$$

where both gamma vectors set γ_0 to zero, or alternatively exclude it (this definition will apply through out the remainder of this chapter unless otherwise stated). The method is readily adapted to branch and bound tree search algorithms (e.g. those of Hendy and Penny 1982, and Turbo-tree Penny and Hendy 1987). Note that OLS on $\hat{\gamma}$ is the same as minimising the sum of absolute deviations for a tree from $\hat{\gamma}^2$ (where each positive entry is squared component wise) (such that any entry in $\hat{\gamma}$ corresponding to an edge in T is not included in this sum). Later in this chapter I discuss how minimising the sum of absolute deviations can be thought of as a type of compatibility tree selection. Hendy and Penny (1993) have shown that branch and bound can find the optimal "closest tree" for up to 20 taxa in short time (a minute or so for "typical" sequences), and we expect nearly identical performance with the OLS criterion (Felsenstein 1993 in the documentation to the branch and bound compatibility program "clique" makes a similar point, although this seems to be contradicted in Kuhner and Felsenstein 1994). With four state conjugations it is proposed that this OLS criterion (and WLS below) would ignore all entries in $\hat{\gamma}$ not mapping directly to a tree edge (the so called "model invariants" of Steel *et al.* 1993c). It may be useful with four or more states, to sum the three or more entries for each potential edge in a tree together (before tree selection) to make the marginal distributions more normal, and to reduce sampling variance.

Application of OLS tree selection to the i.r. Hadamard corrected sequence spectra shown in figure 5.1 (the values on the extreme left of each graph) selects the archaebacterial tree, with a residual sum of squares (SS) equal to $(\gamma_6)^2 + (\gamma_5)^2 \approx 0.03^2 + 0.02^2 = 0.0013$. Edge lengths for this optimal tree are taken directly from the corresponding $\hat{\gamma}$ values shown at the extreme left in figure 5.1.

5.2.4 What is closest tree?

Closest tree (Hendy 1991) identifies the optimal tree as that which minimises the Euclidean distance (E.D.) of tree to the partition (sequence pattern) data,

$$\text{E.D.}^2 = \sum_{i=1}^{2^{t-1}} (\hat{\gamma}_i - \gamma_i(T))^2 + (e+1)(x_T)^2 \qquad (5.2.4\text{-}1)$$

where the summation is over the same entries as that for OLS (i.e. $\hat{\gamma}_i$ not in the tree), but a second tree dependent term $(e+1)(x_T)^2$ is then added on to each tree "score"(e stands for the number of edges in T, i.e. the number of non-zero entries in $\gamma(T)$). The quantity x^T is the sum of all values in $\hat{\gamma}$ not in the tree being examined (γ_0 excluded) divided by one more than the number of edges (e) in that tree. This second factor varies with the number of implied multiple changes in $\hat{\gamma}$ that a given tree does not account for under the model. Negative values can arise in $\hat{\gamma}$, and because these are excluded from being edge weights in $\gamma(T)$, then $\hat{\gamma}(T)$ must lie in the positive quadrant, whereas $\hat{\gamma}$ may lie in any quadrant. When this occurs the optimal tree may well lie on the boundary of two quadrants, i.e. any binary tree with any edge weights set to zero (i.e. a non-binary or unresolved tree). OLS would usually be implemented not picking a negative internal edge weight in the tree, and this is achieved by setting all negative γ_i to zero before tree selection.

It is conjectured then that given the logical condition that $\gamma(T)$ must lie in the positive quadrant, then OLS will always rank binary trees in the same order as closest tree. Specific to this conjecture, is that closest tree will always favour a non-binary tree over a binary tree if one of the edges in the binary tree has to be set to zero because the corresponding γ_i is negative. This insight may be of use as OLS is slightly faster to evaluate than Closest tree. Closest tree was originally described for use on $\hat{\gamma}$, but can equally be applied to observed sequence partitions, $\hat{s}$ or $\hat{f}$.

The closest tree, or smallest Euclidean distance criterion can also be used to infer the edge lengths on the tree (Hendy 1991). It does so by adding the term $1/(e+1)$ (sum of all entries not in that tree, γ_0 excluded) to each entry $\hat{\gamma}_i$ that corresponds to an edge in a specific tree. This redistribution of what are apparently unexplained changes onto the tree(s) of interest is intuitively a reasonable idea (although equal distribution to all edges seems arbitrary). As we will see below, a number of other tree selection criteria (e.g. parsimony) can also be invoked to estimate the edge lengths on a specific tree while taking into account partitions which do not correspond directly to edges in the tree. Some tree selection criteria can also be used to estimate edge lengths (e.g. GLS and ML) given more sophisticated expectations of the sampling distribution of the partition data.

With the i.r. corrected data illustrated in figure 5.1 (extreme left values of figure 5.1), closest tree picks the same tree as OLS, i.e. T_{12} (the archaebacterial tree). However the edge lengths are estimated by adding the term (sum of residuals)/(e+1) to each entry in $\hat{\gamma}$ that corresponds to an edge in this tree. In this case the signals corresponding to the other two possible binary trees (γ_5

and γ_6) is divided up with each edge in the tree T_{12} becoming longer by (0.02 + 0.03)/6 = 0.008. Here this makes almost no difference to the long external edges, but amounts to a nearly 20% increase in apparent support for the internal edge of tree T_{12}. Later we discuss whether this type of redistribution is either biologically appropriate, or optimal.

5.2.5 Weighted Least Squares (WLS)

Tree selection by this general criterion involves searching for the tree that minimises the weighted sum of squares,

$$\text{WLS SS} = \sum_{i=1}^{2^{t-1}} \frac{(\hat{\gamma}_i - \gamma_i(T))^2}{w_i} \qquad (5.2.5\text{-}1)$$

where, in the most general case, w_i is can be any weight estimated for the *i*-th entry in $\hat{\gamma}$. With optimal weighting, WLS tree selection will be more reliable than the previous methods at picking the true tree when the model holds (that is, it has increased statistical efficiency, e.g. Stuart and Ord 1990, p. 1074). If $\hat{\gamma}$ is distributed as multi-variate normal, then the optimal value for w_i is the variance of $\hat{\gamma}_i$ (obtainable from the diagonal of $\mathbf{V}[\hat{\gamma}]$, that is the *ii*-th entry).

A useful step in selecting the optimal WLS tree from gamma is to divided all entries in $\hat{\gamma}$ by the square root of their variance. Such a vector is variance "stabilised". We denote this new vector $\hat{\gamma}_{/se}$, and its *i*-th entry is equal to the number of standard errors (positive or negative) that $\hat{\gamma}_i$ is from zero (this also makes it a particularly useful vector for visual inspection of the corrected sequence data). Algebraic manipulation shows that OLS tree selection from $\hat{\gamma}_{/se}$ is equivalent to WLS tree selection from $\hat{\gamma}$, i.e.

$$WLS(\hat{\gamma}) = \sum_{i=1}^{2^{t-1}-1} \frac{(\hat{\gamma}_i - \gamma_i(T))^2}{V(\hat{\gamma}_i)} = \sum_{i=1}^{2^{t-1}-1} \left(\sqrt{\frac{(\hat{\gamma}_i - \gamma_i(T))^2}{V(\hat{\gamma}_i)}} \right)^2 = \sum_{i=1}^{2^{t-1}-1} \left(\frac{(\hat{\gamma}_i - \gamma_i(T))}{\sqrt{V(\hat{\gamma}_i)}} \right)^2 = OLS(\hat{\gamma}_{/se}) \qquad (5.2.5\text{-}2)$$

Consequently, selection of the optimal WLS tree is just as amenable to branch and bound as is the OLS criterion (this search reduces to finding the minimum sum of absolute deviations when searching for a tree from $(\hat{\gamma}_{/se})^2$, where the squaring is done component-wise). These searches should incur similar run times, although the weightings could results in either faster runs (if there was down weighting signals which were not in the optimal tree) or slower run times (if it tended to more severely down weight signals that were in the optimal tree). In reality (especially when the model does not hold) either situation could arise.

For the i.r. Hadamard conjugation correction of the data listed in the caption of figure 5.1, the standard errors of $\hat{\gamma}$ estimated by the method of section 4.2 are [0.0202, 0.0180, 0.0123, 0.0108, 0.0084, 0.0092, 0.0097] (excluding γ_0). Consequently the vector $\hat{\gamma}_{/se}$ for this transformed data is [10.84, 9.64, **3.71**, 5.41, **2.23**, **3.18**, 4.40] (entries corresponding to possible internal edges are in bold). The counts for all cells in $\hat{\mathbf{s}}$ are substantial (>10) so we are assured of close to normal marginal distributions for all entries in $\hat{\gamma}$. With four taxa (and 2 states) WLS will always

select as optimal, the tree whose internal edge corresponds to the largest of the three eligible entries in $\hat{\gamma}_{/se}$. In this case WLS also selects the archaebacterial tree ($\hat{\gamma}_3$). However, notice how much more similar the support for entries which may correspond to an internal edge has become. Just a glance at the residual WLS SS for each tree (for $T_{12} = 2.23^2 + 3.18^2 = 15.09$, $T_{13} = 23.88$, $T_{14} = 18.74$, $T_{star} = 28.85$) though tells us that the data is statistically very unlikely under the assumed model (similarly, Z test's of the two signals excluded from the tree would both reject their true value was zero, at the $\alpha = 5\%$ level). If the assumptions of multi-variate normality with zero correlations held, then the WLS score would have a χ^2 distribution, with degrees of freedom (d.f.) equal to the number of entries in $\hat{\gamma}$ minus the number of edges in the tree, minus one (e.g. with two states, a binary tree selected, and no other parameters optimised, d.f. = 2^{t-1} - (2t-3) -1).

Applying the WLS criterion to the i.r. Hadamard conjugation corrected transversions from Lake's alignment (see figure 2.7), we would select the eocyte tree as optimal (yet the data would still not fit the model well). OLS in the same instance would still pick the archaebacterial tree. Here is an example where different tree selection criteria applied to $\hat{\gamma}$ result in different trees being picked, although the overriding impression is that the we should be cautious about selecting any tree under this i.r. model. Here we can also see how WLS is in effect making use of the observation in chapter 4, that long edges attracting results in the increased variance of the pattern(s) that group them together (here, this is done by downweighting the signal corresponding to the archaebacterial tree). This feature should increase the robustness of WLS (relative to OLS), with respect to obtaining the correct unweighted tree, when the assumed model is undercorrecting for unobserved changes. This is because the more the long edges attract, the larger the variance of the signal(s) uniting them becomes (Waddell *et al.* 1994, and chapter 4), and the more this signal is down weighted in $\hat{\gamma}_{/se}$. This feature, which suggests robustness, is considered further in the second half of this chapter.

If $\hat{\gamma}$ is distributed multi-variate normal, with zero correlations between entries (that is $\mathbf{V}[\hat{\gamma}]$ is a diagonal matrix), then the WLS tree selection criterion is the maximum likelihood (ML) tree selection criterion from $\hat{\gamma}$. The assumption of independence of entries in $\hat{\gamma}$ will largely hold when rates of change per site are relatively low (again see chapter 4). The assumption that the marginal distribution of $\hat{\gamma}_i$ is normal is well approximated when more than five sites in the sequence have a pattern corresponding to $\hat{s}_i$. Multivariate normality of $\hat{\gamma}$ is often most closely approached when rates of change are moderately high, yet under these conditions correlations are also becoming more prominent (see chapter 4) suggesting a compromise must be met.

In most situations, we expect WLS to be closer to ML tree selection from $\hat{\gamma}$ than is OLS or closest tree, yet there may still be distinct differences. For example, when dealing with sampled data we replace $\mathbf{V}[\hat{\gamma}]$ with $\hat{\mathbf{V}}'[\hat{\gamma}]$, as derived in chapter 4. The use of $\hat{\mathbf{V}}'[\hat{\gamma}]$, rather than $\mathbf{V}[\hat{\gamma}]$, will also result in a difference between weighted least squares and ML tree selection, even if $\hat{\gamma}$ is multi-variate normal, with all entries independent (this last difference can be eliminated by using iterated least squares as is explained in the next section). Indeed estimates of variances usually have notoriously large stochastic errors compared to estimating means, so if using variances as

weighting factors offers an advantage to tree selection with short sequences, will require simulations. An alternative to weighting solely by the variances if their accuracy is in doubt, is to partially weight by them. Such methods are becoming popular in discriminant analysis. Given a certain similarity between this method (and even more so the similarity of GLS and iterated GLS which are introduced next) and ML, the variability of the weighting functions could partly explain why ML tends to do more poorly than other methods with short sequences (e.g. Kuhner and Felsenstein 1994).

A major disadvantage of this method with many sequences is the cost of obtaining the entries v_{ii}. An alternative to calculating $\mathbf{V}[\hat{\gamma}]$ would be to use weights which are similar to the sampling variances of $\hat{\gamma}$. Two possibilities are the sampling variance of the corresponding entries in $\hat{s}$, and using the polynomial approximation to the Hadamard conjugation derived by Szekely *et al.* 1993. We suspect that these approximations will be quite useful, both with regard to increasing the robustness and also the statistical efficiency in comparison to OLS or closest tree.

Lastly, it is worth considering what the weighting should be given to entries in $\hat{\gamma}$ with exponential-like marginal distributions. If all entries had exponential marginal distributions, then the optimal weighting when selecting a tree from $\hat{\gamma}_{/se}$ is to minimise the sum of absolute deviations. Entries in $\hat{\gamma}$ with variances of less than 1/c will tend to have marginal distributions that are more L-shaped than exponential, while variances of greater than 1/c will become more normal until they have an approximately normal marginal distribution when their variance is 5/c or larger. The real danger for tree selection, is not to overweight some of the small entries in $\hat{\gamma}$ which have exponential like marginal distributions. As such a possible solution might be a differential weighting scheme of entries in $\hat{\gamma}_{/se}$, e.g. $\hat{\gamma}_{/wse} = (\hat{\gamma}_{/se})^x$, where $x = ((\text{var } \hat{\gamma}_i \times c)^y)$, where $y = \ln(4)/\ln(5))$ if $(\text{var } \hat{\gamma}_i \times c) \leq 5$, else 2. This weighting would give entries with a normal-like marginal distribution squares weighting, those with an exponential-like marginal distribution linear weighting, and those with very skewed distributions progressively lower weighting. Other weighting functions could be tried, and evaluated with simulations. This weighting, however, should quite effectively downweight the most unreliable entries in $\hat{\gamma}_{/se}$. The use of the sample variance will also tend to negate overweighing of unreliable entries in $\hat{\gamma}_{/se}$, since when these entries are large in $\hat{\gamma}$ this is most likely due to a high count in the corresponding entries in $\hat{s}$, which in turn will increase their estimated variance and see them downweighted more severely. Thus there is a sort of negative-feedback mechanism. Following weighting of $\hat{\gamma}_{/se}$, a tree can then selected from $\hat{\gamma}_{/wse}$ by minimising the sum of absolute deviations (i.e. compatibility tree selection after removing the signs in $\hat{\gamma}_{/wse}$).

5.2.6 Generalised Least Squares (GLS)

It is important to consider not only the variances, but also the correlations of entries, when measuring the deviations of data from its expectation under a certain model. For example, if two entries are near fully positively correlated with one another, then the sum of their deviations from

expectation should count only slightly more than that of an a single independent entry, and not double.

Statistical theory states that a most efficient (i.e. a best asymptotically normal estimator or BAN estimator, Stuart and Ord 1990, chapter 19, pages 707-740) measure for selecting a model from the asymptotically (t fixed, $c \to \infty$) multivariate normal space of $\hat{\gamma}$ is by the minimum GLS criterion. When picking a tree from $\hat{\gamma}$ the sum of squares by the GLS criterion is,

$$\text{GLS SS} = \sum_{i=1}^{2^{t-1}-1} \sum_{j=1}^{2^{t-1}-1} (\hat{\gamma}_i - \gamma_i(T)) v_{ij}^{-1} (\hat{\gamma}_j - \gamma_j(T)) \qquad (5.2.6\text{-}1)$$

$$= (\hat{\gamma} - \gamma(T)) V^{-1} (\hat{\gamma} - \gamma(T)),$$

where $\mathbf{V}^{-1}$ is the matrix inverse of $\mathbf{V}[\hat{\gamma}]$ (γ_0 excluded), v^{-1}_{ij} is the ij-th entry of this matrix, and γ is a column vector. If the correlations (hence covariances) between all entries in $\hat{\gamma}$ are zero (i.e. $v_{ij} = 0$, for $i \neq j$), then this formula reduces to that given earlier for WLS. This and the following five equations may be found in Agresti (1990), particularly pages 460-462. The use of the term GLS vs WLS is not universal; here we mean by GLS a WLS method which uses a statistically efficient estimate of the covariance matrix, and always takes into account correlations between variables. GLS is a very well understood statistical method and when $\mathbf{V}$ is known exactly, then the distance measure $\sqrt{GLS(SS)}$ (also known as the Mahalanobis distance) is the most statistically efficient measure of distance in multivariate normal space (e.g. Manly, 1986, Stuart and Ord 1990 723). When working with a sample, it is usual replace $\mathbf{V}[\hat{\gamma}]$ with its sample estimate $\hat{\mathbf{V}}'[\hat{\gamma}]$, which is derived using the delta method (chapter 4). It is used in many applications, including discriminant analysis, of which tree selection may be thought of as an example. As we have already discovered in chapter 4, when c is sufficiently large, then the distribution of $\hat{\gamma}$ converges to multivariate normal, making this criterion a promising choice for tree selection.

In traditional statistical terms $\hat{\gamma}$ is a vector of expected values with multivariate errors, that is,

$$\hat{\gamma} = \mathbf{X}\gamma(t) + \mathbf{e} \qquad (5.2.6\text{-}2)$$

where $\mathbf{X}$ is a parameter incidence matrix (described below), $\gamma(t)$ is as vector containing just the positive edge weights of the true tree (indexed in ascending order, note the lower case t to indicate this), and $\mathbf{e}$ is a vector of sampling errors described by the variance-covariance matrix, $\mathbf{V}[\hat{\gamma}]$. With the 2-state Hadamard conjugation the edge incidence matrix $\mathbf{X}$ of tree T, has $2^{t-1}-1$ rows and a column for each non-zero edge in the tree, consequently $\mathbf{X} \times \gamma(t) = \gamma(T)$ (γ_0 excluded). We illustrate the form of $\mathbf{X}$ and $\gamma(t)$, with the tree shown in figure 2.1 of this thesis which in vector form gives, $\gamma(T) = [-0.472, 0.2, 0.025, 0.025, 0.2, 0, 0, 0.025]$. Thus,

$$\underset{\mathbf{X}}{\begin{bmatrix} 1 & 0 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 & 0 \\ 0 & 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 1 \end{bmatrix}} \times \underset{\gamma(t)}{\begin{bmatrix} 0.20 \\ 0.025 \\ 0.025 \\ 0.20 \\ 0.025 \end{bmatrix}} = \underset{\gamma(T)}{\begin{bmatrix} 0.20 \\ 0.025 \\ 0.025 \\ 0.20 \\ 0 \\ 0 \\ 0.025 \end{bmatrix}}$$

The matrix **X** for the 4 taxa star tree would be the same as that illustrated above with the third column removed, while $\hat{\gamma}(t)$ would have only 4 entries, lacking the entry corresponding to the internal edge (in the above example the third row). From classical statistical theory it follows that having specified the tree T, the optimal vector $\gamma(t)$ which minimise the GLS sums of squares for a given sample (we shall label these tree optimal estimates $\hat{\gamma}(t)$) is,

$$\hat{\gamma}(t) = (\mathbf{X}^T\mathbf{V}^{-1}\mathbf{X})^{-1}\mathbf{X}^T\mathbf{V}^{-1}\hat{\gamma}, \qquad (5.2.6\text{-}3)$$

where ^{T} defines transpose, and V^{-1} is the inverse of the variance-covariance matrix of the $\hat{\gamma}$ (again see Agresti 1990, or extensive referencing and theoretical development see Stuart and Ord 1990, p. 707-740). Just as usefully the variance-covariance matrix of these optimal edge weights is given as,

$$\mathbf{V}(\hat{\gamma}(t)) = (\mathbf{X}^T\mathbf{V}^{-1}\mathbf{X})^{-1}. \qquad (5.2.6\text{-}4)$$

(which corresponds to the bracketed operations in 5.2.6-3). We may now substitute the formula for $\hat{\gamma}(t)$ into equation 5.2.6-1 and after rearranging terms the optimal GLS SS for a tree is given as,

$$\text{GLS SS} = \hat{\gamma}^T\mathbf{V}^{-1}\hat{\gamma} - \hat{\gamma}(t)^T\mathbf{X}^T\mathbf{V}^{-1}\hat{\gamma}(t) \qquad (5.2.6\text{-}5)$$

Asymptotically (number of taxa fixed, $c \rightarrow \infty$), when the sequences evolved according to the model (the null hypothesis), this residual sum of squares will have a χ^2 distribution with $2^{t-1}-1-p$ degrees of freedom, where p is the number of parameters being optimised (= number of tree edge parameters plus any distribution of rates across sites parameters). If an alternative hypothesis is true, then the residual sum of squares will have a non-central chi-squared distribution (e.g. the SS of a tree other than that which generated the data). The non-centrality parameter may be calculated, and values for the non-central chi-squared distribution looked up in tables or calculated with a computer algorithm (Stuart and Ord 1990, p. 865). This approach extends the analytical method described in section 4.8.2 which assesses how often compatibility picks the wrong tree from $\hat{\gamma}$, to evaluate more sophisticated tree selection methods. These calculations allow the asymptotic power (probability of rejecting the null hypothesis when a specific alternative is true) of the GLS tree selection method to be compared with that of say compatibility or weighted least squares on sequences. A similar approach was taken by Churchill, *et al.* (1992) to estimate the power of ML versus linear invariants. The resulting answers will be accurate for long sequence lengths (say $E[f_i] > 5$, for all *i*), but will become

unreliable as sequence length decreases unless an accurate approximation to the tail of the non-Central chi-squared statistic can be found for the type of sparse frequency vectors found with sequence data.

The other approximation being made, is that the calculations assume that the statistical method has an exact estimate of **V** (see Stuart and Ord p. 724 for some theoretical comments on this matter). As we have already discussed, in actuality a tree building method must use $\hat{\mathbf{V}}'[\hat{\gamma}]$ (or its equivalent), which itself is subject to stochastic errors. Presently simulations would seem the more reliable way to gauge the power of tests and the efficiency of these estimators for the type of data and sequence lengths used in real applications.

A further result from classical GLS statistics is that one can derive the variance-covariance matrix of the residuals. Since the residuals ($\hat{\gamma}$ - $\hat{\gamma}(t)$) are orthogonal to the fit $\hat{\gamma}(t)$, then,

$$\text{Cov}[\hat{\gamma}] = \text{Cov}\{[\hat{\gamma}-\hat{\gamma}(t)]+\hat{\gamma}(t)\} = \text{Cov}\,[\hat{\gamma}-\hat{\gamma}(t)\,] + \text{Cov}\,(\hat{\gamma}(t)) \qquad (5.2.6\text{-}6)$$

so the estimated covariance matrix of the residuals equals $\mathbf{V}[\hat{\gamma}]$ - $\mathbf{V}[\hat{\gamma}(t)]$, where $\mathbf{V}[\hat{\gamma}(t)] = \mathbf{X}(\mathbf{X}^T\mathbf{V}[\hat{\gamma}]^{-1}\mathbf{X})^{-1}\mathbf{X}^T$ (not sure if this is right when we have transformed from **s** to γ).

When dealing with 4-state Hadamard conjugations, then the same equations apply, except that the matrix **X** will now have 4^{t-1} rows, and (2t-3) × 3 columns, with an entry in each one corresponding to the rate of transitions or either type 1 or 2 transversions on a specific edge in the tree.

We next apply GLS tree selection using $\hat{\mathbf{V}}'[\hat{\gamma}]$ to the i.r. transformed data in figure 5.1, with the results shown in table 5.1. Firstly notice how close the GLS sums of squares (table 5.1a) are to those obtained earlier with the WLS method (e.g. 14.00 by GLS vs 15.09 for WLS on the optimal tree). This is due to the correlations of entries in $\hat{\gamma}$ generally being of small to moderate size (the values in $\hat{\mathbf{V}}'[\hat{\gamma}]$ for the real data's i.r. $\hat{\gamma}$ are quite similar to those for the model tree used in chapter 3). Taking into account correlations has, in this instance, made the SS difference between the archaebacterial tree and the eocyte tree larger (i.e. archaebacterial tree slightly more favoured than under WLS), but still provides very clear evidence that the model does not fit ($P < 0.001$). Notice also how the estimated edge weights, $\hat{\gamma}(T)$ (table 5.1b), all show the same biases away from $\hat{\gamma}$ for each of the four trees (just one tree, T_{12}, is shown). This result can be interpreted as the GLS method optimising edge weights to compensate, as far as possible, for the undercorrection that results in none of the possible trees fitting well. These deviations make all external edges longer than they would be if read directly from $\hat{\gamma}$, in an attempt to accommodate evidence of all edges "attracting" more than they should be under the model. On all the binary trees the length of the internal edge shrinks, since the extra length of the external edges explains some of the changes that would otherwise have be explained by an internal edge. The star tree makes its external edges longest, since it does not have the "luxury" of an internal edge to explain any of the bipartition substitutions. We expect that similar effects will be manifest in larger trees, and later we will see the same factors affect edge length estimation with ML

methods. The ranking of the sum of edge lengths by the GLS criteria has some correlation to the SS, although it is not nearly as discriminatory. Rhzetsky and Nei (1992b) discuss sums of edge lengths for OLS, WLS and GLS estimation of edge lengths from pairwise distances alone, and expected a more exact correlation between fit and sum of edge lengths. We suggest that this correlation can break down with violations of the model.

Table 5.1b shows the calculated standard errors of $\hat{\gamma}(T_{12})$, the estimated edge weights which are specific to T_{12}. As with the sums of squares, they are very similar to the errors estimated from $\hat{\mathbf{V}}'[\hat{\gamma}]$. The correlations between estimated edge weights shown in $\hat{\mathbf{C}}'[\hat{\gamma}(T_{12})]$ are also quite like those in $\hat{\mathbf{C}}'[\hat{\gamma}]$ (not shown). In particular we note the moderate negative correlation between the weights of either of the long external edges, and the internal edge (which in this case is a "long edges attract" edge). There is also a moderate negative correlation between the two short external edges, and a moderate positive correlation between the long external edges. All other correlations, as in $\hat{\mathbf{C}}'[\hat{\gamma}]$, are small.

Table 5.1(a). GLS SS and edge weights ($\hat{\gamma}(T)$) for the i.r. $\hat{\gamma}$ of figure 5.1.

		T_{12}	T_{13}	T_{14}	T_{star}
	SS	14.000	22.64	18.10	26.06
i	$\hat{\gamma}$	$\hat{\gamma}(T)$			
1	0.219	0.227	0.244	0.246	0.247
2	0.173	0.181	0.202	0.200	0.203
3	0.046	0.043	0.000	0.000	0.000
4	0.058	0.076	0.069	0.069	0.077
5	0.019	0.000	0.016	0.000	0.000
6	0.029	0.000	0.000	0.026	0.000
7	0.042	0.059	0.056	0.051	0.061
	0.586	0.586	0.587	0.592	0.588

Table 5.1(b). The correlation matrix for the GLS edge weights of T_{12}.

$\hat{\mathbf{C}}'[\hat{\gamma}(T_{12})]$						
			Index			
	$\hat{\gamma}(T_{12})$	s.e.	1	2	3	4	7
1	0.227	0.0201	1.00	0.13	-0.29	0.03	-0.01
2	0.181	0.0179	0.13	1.00	-0.37	0.00	0.01
3	0.043	0.0123	-0.29	-0.37	1.00	-0.02	-0.06
4	0.076	0.0095	0.03	0.00	-0.02	1.00	-0.18
7	0.059	0.0086	-0.01	0.01	-0.06	-0.18	1.00

In chapter 4 we noted that as sequences become longer, functional correlations between sites need not be a serious problem for the convergence of the inferred probabilities of sequence patterns to their true probabilities. Functional correlations can, however, substantially alter the form of $\mathbf{V}[\hat{\mathbf{s}}]$, making variances and covariances larger than they are under the i.i.d. multinomial model. A good example of such non-independence is caused by base pairing in rRNA helices (e.g. Gutell *et al.* 1985). We can estimate these correlations given a tree (weighted in the case of ML), then use parsimony or maximum likelihood to reconstruct ancestral character states, followed by the correlations of changes on the phylogeny for all sites. We could even make these estimates on homologous sequences which shared no part of the evolutionary history of the taxa we wished to analyse, if we assumed that the evolutionary processes operating on sites had not changed since the last common ancestor (e.g. work out correlations of sites in animal rRNA and use this when analysing plant rRNA). GLS is one method that can effectively use this extra information to not only infer trees, but also give estimates of the errors on edge lengths, and the

model as a whole. All that is required is to substitute $\mathbf{V}[\hat{\mathbf{s}}]$ under the multinomial model for the biologically revised matrix and proceed to infer $\mathbf{V}[\gamma]$ and measure fit to each tree. This may not be the optimal method (using information on individual sites coupled with ML may work better) but it should be robust and computationally feasible for up to eight or nine taxa (in 2 character states) working from sequences, and at least up to 12 taxa basing the analysis on distances alone (see below).

5.2.7 Maximum likelihood tree selection from $\hat{\gamma}$

Instead of estimating $\mathbf{V}[\hat{\gamma}]$ from our sample data, we can potentially go one better than standard GLS by employing iterated GLS (Agresti 1990, p 117, 450, 476, Stuart and Ord 1990, p. 1074). Given our initial GLS estimates of a trees edge weights, we can re-estimate $\mathbf{V}[\hat{\gamma}]$ under the assumption that both the tree and mechanism of evolution are correct. To do this we first use the inverse Hadamard conjugation to predict $\mathbf{s}'(T)$ of the weighted tree being considered. (It is also possible predict $\mathbf{V}[\mathbf{r}(T)]$, but due to the speed of the fast Hadamard, it does not offer much advantage, and requires more programming, see chapter 4). Given $\mathbf{s}'(T)$, proceed to infer $\mathbf{V}'^{i}[\hat{\gamma}]$ using the methods developed in chapter 4, by replacing the previous estimate of $\mathbf{s}^{i-1}(T)$, with the newly inferred $\mathbf{s}(T)$ (where the i indicates from the i-th iteration). Given the updated estimate of $\mathbf{V}'^{i}[\hat{\gamma}]$ and the original transformed data, $\hat{\gamma}$, then reestimate the GLS SS score of that particular tree using equation 5.2.6-5 (that is on each cycle substitute $\mathbf{V}'^{i}[\hat{\gamma}]$ (the value for the previous iteration) with $\mathbf{V}'^{i-1}[\hat{\gamma}]$ (its updated value for the edge weights inferred form the previous cycle)). If the value of GLS SS on the i-th iteration is within some value of the previous iteration, stop, else continue the iterations. An alternative to be sure of convergence is to check that the GLS SS for that tree, changes by less than ε on two successive iterations. To evaluate the iterated GLS SS for any other tree (including any contraction of a binary tree), we start over with this new tree's initial optimal GLS edge weighting's, then predict $\mathbf{V}'^{i}[\hat{\gamma}]$ assuming this to be the true tree, and proceed to iterate for this tree until convergence.

Minimising the iterated GLS score is the same as maximising the likelihood when the data is multivariate normally distributed, given that it is necessary to infer $\mathbf{V}$ from the same data that we are fitting a model to (assuming the transformation can be made, i.e. no infinite pathset lengths). Since $\hat{\gamma}$ converges to a multivariate distribution as $c \rightarrow \infty$, *and* since the maximum likelihood point is invariant upon transformation of the original data (Stuart and Ord 1990, p. 685), then it follows that as c increases this method converges to maximum likelihood (ML) on sequences (Felsenstein 1981, and described below). If c becomes large enough, then under a Markov tree model of sequence evolution, there must exist one global GLS SS optimum for any given tree if we can estimate $\mathbf{V}$ consistently. This in turn implies but one optimum per tree for ML on sequences (for Hadamard models, but probably also others) if the model is correct (e.g. Stuart and Ord 1990, p. 1075). If the assumed model is not correct, but all entries in $\hat{\mathbf{s}}$ have positive values, then if iterated GLS can be shown to have a single minimum asymptotically, so too must ML. I have been told that there are proofs of the uniqueness of iterated GLS solutions (G. Arnold

pers. comm.) but citing one has proved elusive (Stuart and Ord 1990, p. 1075 mention what appear to be partial proofs).

The required conditions, if such a proof exists, probably include that the true **V** must be positive definite and non-singular, and that all pathsets are of finite length. A Markov tree model of evolution does imply that every entry in **s** will be non-zero as long as all external edges are non-zero, since then there must be a finite positve probability of every pattern occurring (if any entry in **s** is non-zero then **V** will not be of full rank, and there need no longer be only one solution, i.e. optima per tree). Others have claimed proofs of a single optimum for ML (e.g. Fukami-Kobayashi and Tateno 1991, and Tillier 1994), but Steel (1994b) has disproved their claims with a counter example. The difference is that our conjectures rely upon certain asymptotic regularity conditions being meet (e.g. $c \rightarrow \infty$, finite pathset lengths, etc.) which usually will not apply with real data. However, study of them may reveal the type of realistic situation where multiple ML optima per tree are most likely to exist.

This relationship between iterated GLS and ML also suggests a lower bound on the number of iterations necessary to converge to the optimum GLS or ML point, with a certain accuracy. When using a Newton type numerical search method in a multivariate normal space (or any smooth continuous function), then the best convergence that can be achieved is quadratic convergence (e.g. Minoux 1986, where quadratic means the error e from the true limit of the GLS iterations will be only e^2 on the next cycle), when already close to the maximum likelihood point. In reality we will not be assured of multivariate normal conditions (but certainly a smooth likelihood function, with respect to changes in the model parameters), so hopefully such analyses will converge at a considerable portion of this optimal rate.

To illustrate this criterion, iterated GLS was run for the four possible 4-taxon trees using the i.r. transformed data of figure 5.1. Table 5.2 shows the results of iterated GLS on $\hat{\gamma}$ for tree T_{12} (the archaebacterial tree). After the first 2 iterations convergence of the GLS SS is rapid. In this table we have also shown the fit in **s** for each iteration as measured by both the lnLR statistic (commonly quoted as $G^2 = 2\text{lnLR}$) and also Pearson's X^2 statistic. The GLS SS actually comes very close to the G^2 and X^2 optima on the first iteration (for G^2 better than 2 decimal places), and then drifts slightly away. The same features of a slightly more rapid and smother series of G^2 and X^2 statistics through the iterations, as well as the greatest similarity of the three statistics (SS, G^2 and X^2) after the first iteration, was repeated for the other three trees (results not shown). The further the divergence of model from data, the greater the discrepancy between G^2 (or X^2) and the iterated GLS SS optima became, reaching a maximum difference of 0.5 lnLR units for the highly non-optimal star tree. This behaviour is understandable as the data is not truly multivariate normal, and it is also known that increasing i (the number of iterations) does not always lead to better estimators (Stuart and Ord 1990, 1075).

From table 5.2a it appears that the iterated GLS SS statistic is less sensitive than either of the other two statistics (G^2 and X^2) in detecting departures from the model. This observation is possibly connected with the trend of GLS underestimating the long edges slightly, and frequently

overestimating the short edges relative to the edge lengths inferred by either G^2 or X^2. Table 5.2b shows the optimal iterated GLS edge weights for T_{12}. The internal edge lengths estimated by iterated GLS on the different trees, could be either slightly longer or shorter than their counterparts estimated by the minimum X^2 or G^2 criteria. The differences in edge length varied from almost nothing on T_{12}, becoming more noticeable on the worst fitting star tree (edge 4 = 0.0771 by iterated GLS SS, but 0.0841 by maximum likelihood) though with no apparent trend as to which edges would be effected most. These differences are minor and the differences between GLS SS and G^2 (but not X^2, since this statistic has different properties) are expected to disappear as c becomes longer and convergence to multivariate normality increases (part of the difference may also be due to the delta method of estimating **V**, which also becomes more accurate as *c* increases).

Estimated standard errors by the iterated GLS criterion (table 5.2) are larger than those estimated by GLS without iteration (table 5.1). This is not an attribute of the criteria of iterated GLS *per se*, but rather a re-alignment of the covariance matrix **V** with the predicted values in $\mathbf{s}^i(T)$, which have resulted in some larger **r** values (i.e. longer pathsets on the tree, which now has longer edges). From the point of view of tree selection, possibly the most important change is that while the internal edge (entry 3) itself became smaller its variance increased slightly. This is understandable in light of the observation that longer external edges create greater variance of internal edges (Waddell *et al.* 1994, section 4.2.5). Churchill *et al.* (1992) state similar results in the context of discussing the power (ability to discriminate one tree from another) of maximum likelihood estimates of phylogeny. The lesson is to keep external edges as short as possible by either choosing more slowly evolving taxa, or by including more taxa branching deeply so as to reduce "long edge" attract type effects. An alternative approach which deserves study is the use of ancestral sequences, which may cut considerable length off external edges.

Table 5.2(a). GLS iterations for T_{12}.

	GLS SS	G^2	X^2
it. no.	14.00359	18.25770	21.73360
1	18.47272	15.47231	16.68747
2	13.87440	15.48070	16.69354
3	13.87528	15.48079	16.69325
4	13.87494	15.48081	16.69319
5	13.87488	15.48081	16.69318
6	13.87486	15.48081	16.69318
7	13.87486	15.48081	16.69318

Note: G^2 and X^2 were both measured at **s** level after each iteration, with $\mathbf{s}^i(T)$ predicted with GLS edge weights.

Table 5.2(b). The final covariance matrix of estimated edge weights, $\hat{\gamma}(T)$.

edge	lengths		$C[\hat{\gamma}(T)]$	Index			
Index	$\hat{\gamma}(T_{12})$	s.e.	1	2	3	4	7
1	0.2446	0.0210	1.00	0.12	-0.32	0.00	0.00
2	0.1960	0.0189	0.12	1.00	-0.40	0.00	0.00
3	0.0336	0.0127	-0.32	-0.40	1.00	-0.02	-0.07
4	0.0830	0.0104	0.00	0.00	-0.02	1.00	-0.25
7	0.0657	0.0095	0.00	0.00	-0.07	-0.25	1.00

The iterated GLS fit of each of the 4 trees is shown in table 5.3. While the fit statistic has become smaller for each tree, it is still significantly bad in all cases. The rank order of the trees remains the same, but the halobacteria tree is now nearly indistinguishable in fit from the star

tree. The general trend was for edge lengths to increase over their previous values under GLS, however this increase was non-uniform, with some edges decreasing slightly in length while many others increased in length by 5 to 20%, with some of the shorter edges increasing by over 50% (e.g. e_7 of T_{12}).

Table 5.3 Iterated GLS results for the i.r. transformed $\hat{\gamma}$ of figure 5.1

index	$\hat{\gamma}(T_{12})$	$\hat{\gamma}(T_{13})$	$\hat{\gamma}(T_{14})$	$\hat{\gamma}(T_{star})$
1	0.244	0.271	0.268	0.272
2	0.196	0.224	0.220	0.225
3	0.033	0.000	0.000	0.000
4	0.083	0.082	0.076	0.084
5	0.000	0.003	0.000	0.000
6	0.000	0.000	0.017	0.000
7	0.066	0.066	0.058	0.068

5.2.8 How many likelihood optima per tree?

Steel (1994b) has recently shown that there may be more than one likelihood optima per tree, and demonstrated this with a 4-taxon 2-state example. Since we have shown that asymptotically our data becomes multivariate normal, so GLS has but one optima per tree (given certain constraints) and that iterated GLS converges to ML, then are these findings contradictory to those of Steel? The answer is no, but rather these findings should help answer Steel's (1994b) question of how general his finding is.

Steel's (1994b) example lies on the boundary of the likelihood space. It is recognised in statistical theory that such multiple optima can exist, but in many cases they exist only on the boundaries of the likelihood space (e.g. Stuart and Ord 1990, p.649) and so can be readily identified. It remains to be shown that all multiple optima for maximum likelihood on sequences are not of this form (including when invariant sites are taken into account by the likelihood model, since the only way suggested so far to move the minima away from the boundary of the parameter space (Mike Steel pers. comm.) is by adding invariant sites to the data).

I believe it is important that all Mike Steel's examples (1994b, and pers comm.) rely upon some patterns not being expressed in the data. When this occurs the covariance matrix estimated via the delta method does not have full rank (i.e. some rows and columns are linear combinations of others). If the data comes from under the assumed model, then asymptotically, $\mathbf{V} = \mathbf{V}'[\hat{\gamma}]$, and there can be but one GLS optima and one ML optima (as long as no edge in the tree is infinitely long). This is because under these circumstances (given that all external edges in the tree have a positive weight) all entries in $\mathbf{s}(T)$ must have a positve probability of occurrence (so by the law of large numbers there will be many of each pattern $c \rightarrow \infty$); and thus $\mathbf{V}$ will be positive definite and of full rank (γ_0 excluded).

It is important also to consider the limit when the data does not fit the model, but still comes from a Markov model (perhaps a mixed Markov model) such that all entries in $\mathbf{s}(T)$ will

asymptotically correspond to many observed sequence patterns. The covariance matrix for each s(T) should still be unique and of full rank **V**. The question is then: can there be two (or more) weighted but otherwise identical tree's that can generate **V** matrices which will each generate local optima under iterated GLS? We suspect the answer is no, but do not have a proof at the moment. Conjecturing one step further we suspect that ML will always have one optima per tree if all possible sequence patterns are observed in the data, and solutions do not lie on the boundary of the parameter space (e.g. no. implied edge lengths are infinite) (but again do not have any proof of this).

Another useful angle on Steel's (1994b) finding is, does this same problem occur for the closely related method of minimum X^2 tree selection? (See Read and Cressie 1988 on the relationship between ML and X^2 fitting). The answer appears to be no. For Steel's (1994b) example there seems to be one X^2 optima per tree and this finds the optima as an average, rather than as an 'either or' situation with branch lengths. Part of the reason for ML having multiple optima is that it "ignores" patterns which are not observed in the data, when estimating fit of data to model.

5.2.9 Comparing GLS on sequences with GLS on distances

Minimum GLS SS has been suggested as a method of tree selection from distances since Cavalli-Sforza and Edwards (1967). About the only examples of its implementation are in Hasegawa *et al.* (1985) where they used it as an approximation to maximum likelihood (which at the time was deemed too computationally expensive), and Bulmer (1990) who redescribed such methods in terms of traditional statistical linear models (neither considered iterated GLS). Generally a distance matrix will become multivariate normal much more quickly than $\hat{\gamma}$, since it has fewer entries, all of which are estimated by the changes on at least two edges in the true tree (and if these count five or more changes the approximation becomes reasonable). Thus the methods assumptions will tend to be met with shorter sequences, and / or more taxa than GLS on $\hat{\gamma}$, since convergence to multi-variate normality is quicker.

Consequently iterated GLS tree selection may converge quite quickly (with increasing c) to maximum likelihood phylogeny estimation when the data is limited to pairwise distances (for whatever reason). Iterative GLS is presently limited to methods which can infer the theoretical variance-covariance matrix of a weighted tree, and a model of sequence evolution. At present the only such methods are the 2-state model presented here, and an analogous 4-state model discussed by Bulmer (1991a)(the equal input model). Closed form expressions of the covariance between estimated distances should exist via the delta method for most transformations (including the general time reversible distance, and the LogDet). Alternatively non-iterated GLS can be performed on most distance data, using the bootstrap (jackknife, or other resampling method) to infer the covariance matrix if an analytic expression is unknown.

Having a well understood statistical method being applied to both transformed sequences and distances, allows an opportunity to answer Penny's (1982) query of what fundamental difference's are there between these two main approaches to tree selection. With four taxa and 2

states there are only 6 pairwise distances, while $\hat{\gamma}$ has one more degree of freedom with 7 independent entries. Consequently we should expect GLS tree selection from $\hat{\gamma}$ to be more discriminating, dealing as it does with more finely categorised data.

Below in table 5.1 we show the results of applying GLS to the i.r. transformed data distances from the data in figure 5.1. As with the other methods, the archaebacterial tree is best fitting. However a number of things stand out as different to GLS on $\hat{\gamma}$. Firstly, the smallish residual of T_{13} is due to the method inferring a negative internal edge length. This is not a problem in itself (since for interpretability we simply collapse to the star tree), but is distinct from all the sequence based methods examined here. Even with T_{13} relegated to synonymy with the star tree, a striking feature of the GLS SS on just pairwise distances is how good the fit of each tree to data appears (and also how similar the eocyte trees fit is to that of the star tree). The residual SS is expected to be distributed close to a chi-square distribution with 1 degree of freedom for the binary trees, and 2 degrees of freedom for the star tree with this much data. The SS under the non-iterated GLS distance method would not reject the null model (α = 5%) that the data fits adequately, yet the same test using GLS on $\hat{\gamma}$ clearly did ($P < 0.001$). Later, we find more evidence to suggest that GLS on distances is not as sensitive at detecting lack of fit of data to model as GLS methods based on sequence patterns.

The optimal edge weights for GLS on distances (table 5.4) take values intermediate between those of GLS on $\hat{\gamma}$ and iterated GLS on $\hat{\gamma}$. One interpretation of this is that they have, due to the greater freedom of fitting distances onto a tree, been able to adopt values closer to being additive on the tree. That is, fitting distances to a tree is less constrained than fitting a tree about three near orthogonal partitions, each of which can fit exactly only one of the three trees, and will yield a substantial SS if the tree being fitted does not match the trees internal edge. If this interpretation is correct, then we would expect the standard errors of edge lengths fitted by GLS on distances to be larger than those when fitted via GLS on $\hat{\gamma}$, and this is indeed observed (contrast s.e. entries in table 5.1b with those in 5.4b). With GLS on distances, the weights on all possible internal edges have shrunk; in this instance there is a trend towards a more star like tree with respect to the sequence based methods. Consistent with predictions, the correlation matrix of inferred edge weights for T_{12} (table 5.4b) shows uniformly higher correlations than those of GLS on $\hat{\gamma}$, but importantly shows the same overall structure (the higher correlations are because the optimal tree is suggesting some longer path lengths).

Thus for a partial answer to Penny's 1982 question of what fundamental differences are there between tree selection on distances vs sequences we suggest:

(A) That distance based methods are statistically less powerful at detecting at least some deviations of the data from the models.

(B) Distance based methods higher variances on estimated edge weights.

Both of these features are important when analysing data, and the second tends to suggest that tree selection using sequence pattern based methods will be more statistically efficient as the

sequence length becomes large. It also suggests that methods such as ML should do better at estimating relative divergence times of nodes, which is a major use of phylogenetic trees in studying both molecular and organism evolution, as well as biogeography, anthropology etc. With short sequences, distances may do better as there are fewer elements being estimated in **V** for example.

Table 5.4(a). Tree selection by GLS applied to transformed pairwise distances, compared to $\hat{\gamma}$.

		T_{12}	T_{13}	T_{14}	T_{star}
	SS	0.769	1.200	3.349	3.384
i	$\hat{\gamma}$	Optimal edge weights by GLS on the distance matrix			
1	0.219	0.242	0.259	0.253	0.254
2	0.173	0.197	0.213	0.209	0.210
3	0.046	0.022	0.000	0.000	0.000
4	0.058	0.081	0.092	0.082	0.083
5	0.019	0.000	-0.016	0.000	0.000
6	0.029	0.000	0.000	0.002	0.000
7	0.042	0.067	0.076	0.068	0.069

Table 5.4(b). Correlation matrix of the edge weights of T_{12} inferred by GLS on distances.

$\hat{C}'[(T_{12})]$							
			Index				
	(T_{12})	s.e.	1	2	3	4	7
1	0.2422	0.0205	1.00	0.17	-0.34	0.06	0.04
2	0.1971	0.0184	0.17	1.00	-0.43	0.04	0.07
3	0.0219	0.0135	-0.34	-0.43	1.00	-0.09	-0.15
4	0.0813	0.0097	0.06	0.04	-0.09	1.00	-0.13
7	0.0667	0.0089	0.04	0.07	-0.15	-0.13	1.00

We now round off the discussion of GLS methods in general. Mitigating against using GLS (particularly iterated GLS) for tree selection is its expense to compute. For the Hadamard conjugation models described in chapter 2, the cost of one iteration of GLS is of the general order $O(n^3)$ (where $n = 2^{t-1}$), due primarily to the need to invert the variance-covariance matrix. This reason for shying away from iterated GLS is in contrast to what has been a prevailing theme in more traditional statistics where iterated GLS has been used to find ML solutions, since computer packages could do this readily when the ML solution had not been programmed (Agresti 1990). Another obstacle to using GLS for tree selection from $\hat{\gamma}$ is that many (or even most) entries in $\hat{\gamma}$ may not have anything like multivariate normal marginal distributions (chapter 3). This may not be too much of a problem for tree selection *per se*, as the method is known to be quite robust to non-normality, but it will present a problem if **V** is not of full rank, due to some sequence patterns being unobserved in a sample. There is no easy way around this problem; standard methods such as putting small constant values into the unobserved cells are often arbitrary and can produce undesirable effects (a more reasonable alternative may be predict the star trees edge weights by some other criteria (say ML or straight from $\hat{\gamma}$), and fill in the missing values in $\hat{\mathbf{s}}$ with those of this **s**(T) (and adjusting so that the whole **s** vector still sums to 1). Hopefully using the star tree will minimise any bias to select particular trees due to partial conditioning of **V** on a specific tree.

A further possible disincentive to using GLS on "typical" sequence lengths comes from recent work on discriminant analysis via simulations, which suggest that GLS may often be less effective than Euclidean distance or OLS in allocating samples to the correct group. This reversal

of asymptotic statistical efficiency appears to be due to the potential for greater error in estimating higher order moments (such as **V**) relative to estimating just their means (the same sort of thing happens in discriminant analysis, Charles Lawoko pers comm.). Whether this is likely to be the case for "typical" phylogenetic analyses will require simulations which are beyond the scope of this thesis. These limitations suggest that GLS methods may have their greatest appeal in testing fit of tree to data in order to test the model, estimate edge lengths, and / or identify a confidence set of trees. Chapter 3 has additionally identified optimising fit of distances to a tree by allowing unequal rates of change across sites as an important usage of GLS on distances.

5.2.10 Statistical properties of compatibility and parsimony applied to gamma.

Here we briefly review what is known of the properties of these well known tree selection criteria. Beginning with compatibility, the tree this method will chose from $\hat{\gamma}$ is that with the maximum sum of edge weights (where edge weights are taken as the value of $\hat{\gamma}_i$). Or alternatively compatibility searches across trees with the aim of finding the tree with the lowest sum of absolute deviations,

$$\text{SAD} = \sum_{i \in T} |\gamma_i|, \ \gamma_0 \text{ excluded,} \qquad (5.2.10\text{-}1)$$

and with the additional constraint that if i is a member of T and negative, then the sum of such $\hat{\gamma}_i$ is added to the sum of 5.2.1.-1 (this prevents the resolution of edges in favour of the worst supported). One situation where this becomes a maximum likelihood method, is when entries in $\hat{\gamma}$ are uncorrelated, and each one has an exponential distribution with equal variances. The closest $\hat{\gamma}$ is expected to come to this situation is at low rates of change and / or very short sequences, when the marginal distribution of edges becomes binomial with expected values in the corresponding entry in **s** being about 1 (thus these binomial distributions take on a form like a discrete exponential distribution). Felsenstein (1981b) points out some other situations where compatibility is closely related to a maximum likelihood solution, notably when a few characters evolve at a high rate of change, while the remainder have very low rates (but we do not know which are which). Kuhner and Felsenstein (1994) observed compatibility to do moderately well in this circumstance with simulations. As mentioned earlier in section 4.8.2 Charleston (1994), ran simulations where the sequences of two trees are combined. He attributed the standout success of closest tree and compatibility with the Hadamard conjugation, to the Hadamard conjugation: we attribute it mostly likely to compatibility and compatibility-like tree selection. The simulations most probably failed to show this clearly since the necessary full set of comparisons were not made by Charleston (1994).

Parsimony is a very interesting criteria because it uses a weighting function based on the tree being evaluated. For example with say 10 taxa and 2 states, then depending on the tree being evaluated, an entry $\hat{\gamma}_i$ which does not correspond to an edge in that tree could be weighted by a factor of between two and five times. Parsimony then is a weighting function which is finely tuned to the hierarchical structure of a tree, and severely punishes (in terms of the weighting

function it applies) any tree which does not have partitions close to the partitions being observed in the data. It is not clear how much robustness this type of weighting offers, but it would be expected to accelerate the rate of convergence. This acceleration can be increased in speed by essentially applying a double weighting, something first implemented by Farris (1969). In this case one selects a tree and then reweights partitions in the data by some function of the number of steps such a character must have had on the best tree. A similar effect could be gained by applying a weighting to the cost of parsimony changes. For example, when searching a tree from **s** by parsimony, let the cost to the length of the tree being evaluated be (parsimony length)^x, where the value of x could be 0, 1, 2, ... (or any real number, or real valued function). If the value of x is 0, we have compatibility, if it is 1, we have standard parsimony, if it is 2 we have a double parsimony, etc. Instead of raising to a power, Felsenstein (1981b) suggests the cost should be ln(parsimony length), since this function approximates the likelihood. There is much to understand about these different methods, and when one will work better than another.

It is of course also possible to apply either compatibility or parsimony to $\hat{\gamma}_{/se}$. Compatibility applied to $\hat{\gamma}_{/se}$ aims to minimise,

$$\text{tree length} = \sum_{i=1}^{2^{t-1}} \frac{|(\hat{\gamma}_i - \gamma_i(T))|}{\sqrt{w_i}}, \qquad (5.2.10\text{-}1)$$

where w_i is the variance of the i-th entry in $\hat{\gamma}$. This criteria is similar to weighted least squares (eq. 5.2.5-2) and would be the ML method if entries in $\hat{\gamma}$ were independent and had marginal distributions that were exponential. Parsimony, in turn, is applying a tree dynamic weighting function to the cost of linear deviations. It is not clear if this will be an especially useful combination under the model, since the Hadamard transform, followed by weighting to standardise variance, is expected to have already downweighted patterns due to parallelisms and convergences. However, if the assumptions of the conjugation and weighting are not correct, then the dynamic weighting of parsimony could well help in quicker convergence and perhaps some degree of robustness. Like all the other weighting methods, much careful study is needed before any argument for the superiority of these types of method can be accepted. And of course many other mix and match methods of tree selection can also be conjured up. The following one does seem to have some interesting properties.

5.2.11 Non-iterated likelihood and non-iterated X^2.

An interesting alternative to iterated likelihood is to evaluate likelihood without iterating edge lengths. One way of doing this is to take the edge weights in $\hat{\gamma}$ as the given edge weights, of a tree T_n, then use a Hadamard conjugation to predict $\mathbf{s}(T_n)$. We can then easily measure the fit of $\mathbf{s}(T_n)$ to the observed data $\hat{\mathbf{s}}$. One measure of fit, which we will see later has surprising robustness, is selection of the tree with the minimal G^2 statistic (which is the same as the ML tree, as described more fully in section 5.3.1). An alternative method of tree selection which appears to do even better, is to find the $\mathbf{s}(T_n)$ which has the lowest Pearson X^2 statistic (again defined in the next section, and this too has some surprising robustness properties). Another

potentially useful application of likelihood evaluation of trees without iterative optimisation is to help infer distributions of rates across sites (and associated shape parameters). Fully optimised ML is probably preferable for this purpose, but computationally this approach would be much quicker. Either approach will often require simulations in order to calibrate any secondary statistical test one might given the sparseness problems the data often poses (see chapter 6).

There is no reason that this criteria of non-iterated likelihood could not be applied secondarily to other methods which select a weighted tree. For example one could find the tree with edge weights estimated by OLS or WLS applied to just the distances which has the highest likelihood (for many models likelihood = the G^2 statistic, would be preferable to X^2 because with large trees, the later method must involve estimating the expected probabilities, or likelihoods, of the exponentially increasing number of possible sequence patterns). Such a method could be quick, and the results of simulations might well tell us something useful about the nature of tree selection. It will be especially interesting if the statistical efficiency or robustness was on a par with iterated ML methods, which have been shown to be quite reasonable themselves (e.g. Kuhner and Felsenstein 1994, references discussed therein, and later in this chapter).

5.2.12 Statistically efficient criteria to choose amongst the best trees

Unfortunately, back conjugation, and iterated ML methods often require a lot more computation than direct selection from $\hat{\gamma}$. One possible solution is applying the most sensitive, and perhaps robust methods, to a set of reasonable trees selected by a quick but still robust method e.g. use maximum likelihood to rank best 100 WLS from $\hat{\gamma}$ trees. This sort of mixed method deserves more study. Apart from just strict tree selection, relative rankings of the best trees by different methods are a useful descriptive statistic. One for example could use these ranks to look for features that might indicate possible inconsistency (e.g. quite different ranks with different methods, something which can be evaluated with non-parametric tests). We encountered a number of instances where relative ranks appeared to differ by a significant amount between methods in chapter 3 (especially interesting were the differences in the transversion trees via parsimony applied to the observed sequences, and neighbor joining applied to the Poisson transformed transversion distances. Other differences observed in section 3.7.3 regarded the placement of the nematode and the *Crithidia* sequence). Similar issues have independently been considered by DeBry and Able (1995). A critical outstanding issue is whether the correlation of the rank of quite suboptimal trees, is a good predictor of the correlation of the near very near optimal trees.

An extension would be to consider a secondary weighting by rank, and a decision theory method to choose a best tree. Klenk *et al.* (1994), used such a method (although with debatable effectiveness as work with P.J. Lockhart has shown the alignment to be the major problem with the RNA polymerase data they used). A promising reason for using some sort of discriminant analysis on ranks of trees by different methods comes from the simulations of Kim (1993). These simulations showed good increases in statistical efficiency of tree selection under a simple model when the estimate of the true tree was taken as the tree selected by the majority of three different

methods. Of course in real situations confidence in the consistency of different methods should preclude taking a median (e.g. in chapter 3 most nearly all current methods suggest that *Giardia* branched more deeply than Microsporidia, yet this is most probably an artifact in all these cases).

5.2.13 Using these methods to select a consensus tree from bootstrap proportions.

There are some interesting similarities between tree selection from $\hat{\gamma}$ (with 2-state data especially) versus tree selection from bootstrap proportions. For example, the majority rules consensus approach can be though of as tree selection from $\hat{\gamma}$, limiting entries to those signals with bootstrap partition proportions of greater than 50%, followed by tree selection using compatibility. Similarly, the Nelson consensus tree method can be thought of as tree selection from all partitions in a set of trees, using compatibility (R.D. Page, pers. comm..).

Bootstrap proportions of tree edges tend to have counts of more than 20 for all edges in the tree (even with only 100 resamplings), thus the marginal distributions of any edge likely to be selected in the tree can be taken as close to normal. In addition while the frequency of occurrence of edges will be close to independent if they are in very different parts of the tree, this need not apply if edges are close to each other in the tree. In this case bootstrap proportions of support may be correlated (either positively or negatively). Accordingly, it would seem that a method such as Generalised least squares might be useful in selecting a consensus tree, since it would be taking into account the locally strong non-independence of edge selection via the covariance matrix of these tree "signals" (and of course assuming that we expect all errors to be random errors about one correct tree). Even without using this method, on occasion it would be interesting to study the correlations of the selection of edges, looking especially for distinctive features.

5.3 ML TREE SELECTION FROM THE OBSERVED SEQUENCES

Maximum likelihood phylogenetic analysis on sequences is typically done using the algorithm of Felsenstein (1981a). In this chapter we illustrate how Hadamard conjugations can be used to make the necessary calculations. Using this approach combined with the extended Hadamard conjugations discussed in chapter 2, we have implemented maximum likelihood (ML) tree selection under models which allow continuous distributions of rates across sites (e.g see Waddell and Penny 1995).

The steps in ML tree selection are conceptually simple. Given a weighted tree and a mechanism of character change, then for a single site pattern, calculate its probability of occurring on that tree model. Do likewise for all the other sites, and multiply these probabilities together to get the overall likelihood of obtaining the observed data under the hypothesised model. One must then heuristically search for the optimal set of edge weights (and any other

variable parameters, e.g. tr/tv ratio) to maximise the likelihood (and find the tree model which fits best). Felsenstein (1981a) initially used an procedure called the EM algorithm, and more recently has found a direct search method quicker (Felsenstein 1993). Recently it has been shown that there can be more than one likelihood optima per tree (Steel 1994b), and this problem must also be considered a possible complication to the optimisation procedure. The overall problem of finding the maximum likelihood point (as we will illustrate later), fits naturally into a more traditional numerical optimisation framework, with its large body of theoretical and simulation results. Following optimisation of likelihood on one tree, comes the problem of searching across different trees to find the weighted tree with the highest likelihood. In order to extend the maximum likelihood tree selection criteria to more general models of sequence evolution, the critical step is to calculate the likelihood of the data under the model. If sites evolve independently, then the critical step reduces to being able to estimate the probability of a single site pattern under the model (the overall likelihood is just the product of these site likelihoods).

In this section we use the results of chapter 2 where we where predicting the probabilities of site patterns, to illustrate finding trees by ML. Following this, we explore finding trees by ML when there is a distribution of rates across sites. If the true distribution of rates across sites is continuous we consider the use of discrete approximations (step functions, or crude numerical integrations) in order to speed up the process. It is important to consider the benefits of not forcing the distribution of rates across sites to a predetermined distribution, as a way to learn more about molecular evolution. The question of estimating likelihoods under models where there is a non-stationary distribution of rates across sites is explored, and a way of calculating likelihoods in this situation is described. The covarion model (Fitch and Markowitz 1970) is one example of this type of process, an issue considered also in appendix 2.3 and chapter 3. Another realistic situation of sites changing their relative rates occurs with a functional to pseudogene transition. This is important to consider since trees containing both functional and pseudogenes are often particular interest (e.g. using functional genes as the outgroup to pseudogenes, as is the case with the psi-eta region of primate nuclear DNA).

Part of this section also considers the important question of whether branch and bound methods can offer a significant advantage when searching tree space for the maximum likelihood tree. We focus on the all important problem of obtaining a tight bound. We briefly look at numerical methods for finding the ML of a specific tree, and consider the role of the Hessian matrix in doing this.

5.3.1 Calculating likelihood via Hadamard conjugations

Here we look at the steps involved in fitting a tree model to sequence data, and then taking the optimal tree from the tree model which best fits the data. We begin by using the log likelihood ratio (or G^2 statistic) as our criteria of fit between $\hat{\mathbf{s}}$ and $\mathbf{s}(T)$, then show how this statistic is directly related to the likelihood statistic commonly used in phylogenetics (e.g. by Felsenstein 1981a).

When working with Hadamard conjugations it is easy to obtain the probabilities of all possible sequence patterns, **s**(T), under specific models. It is straight forward to then calculate the natural logarithm of the likelihood ratio statistic (lnLR) of a set of real data as,

$$\ln LR = \sum_{i=0}^{2^{t-1}-1} \hat{f}_i \ln\left(\frac{\hat{f}_i}{f_i(T)}\right), \qquad (5.3.1\text{-}1)$$

where $\hat{f}_i$ is the observed and $f_i(T)$ (= c × s_i(T)) the inferred frequency of a sequence pattern (and it is taken that when $\hat{f}_i = 0$, then the function to the left of the summation sign takes value zero). Zero is the appropriate value for the contribution of an unobserved pattern to the overall likelihood function. One way of understanding why zero is the appropriate value (rather than undefined) is to consider what happens to this term per cell if the observed count is allowed to behave as a real valued variable (and f_i(T) is effectively fixed in its range). In this case we have two opposite trends occurring: $\hat{f}_i$ is heading towards zero from above at a linear rate, as is the term $(\hat{f}_i \,/\, f_i(T))$. The ln of the term in brackets is heading towards negative infinity at a decelerating rate (relative to linear), so overall the whole expression is heading towards zero from below and is considered to reach zero in the limit as $\hat{f}_i \to 0$. We give the summation up to 2^{t-1}-1 as the index of the last pattern under the 2-state Poisson model. If working under the general i.i.d. 2 state model we have 2^t possible patterns (and **s**(T) must be calculated by some other method than a Hadamard conjugation). As we have already noted if working under the generalised Kimura 3ST model, there are 4^{t-1} patterns, while the most general i.i.d. 4 state model has 4^t distinct pattern probabilities.

The lnLR statistic is a measure of fit of data to model, specifically the logarithm of the ratio of the likelihood of the data under the tree model to the likelihood of the data under the unconstrained multinomial model. This statistic will be zero only when $\hat{\mathbf{s}}$ and **s**(T) are identical, and becomes larger the worse the fit between them. Accordingly, we seek to find a tree model with vector **s**(T) which minimises this statistic. The G^2 test statistic (e.g. Read and Cressie 1988) is sometimes also called the G statistic e.g. Sokal and Rohlf (1981, p. 721), while Stuart and Ord (1990, p. 1160) refer to it as -2 times the LR statistic. It is an alternative to the well known chi-square goodness-of-fit statistic (X^2) (which it is very similar to, and which we also use in model fitting and tree selection). When using the X^2 statistic to select a tree model, the aim is to minimise the statistic,

$$X^2 = \sum_{i=0}^{2^{t-1}-1} \left(\frac{\hat{f}_i - f_i(T)}{f_i(T)}\right)^2, \qquad (5.3.1\text{-}1)$$

where $\hat{f}_i$ is the observed and $f_i(T)$ (= c × s_i(T)) the expected frequency of a sequence pattern.

Table 5.5 Optimal edge weights for the data in figure 5.1 fitted to the i.r. 2-state Poisson model by minimising the G^2 (= 2 × lnLR) or X^2 statistic.

	T_{12}		T_{13}		T_{14}		T_{star}	
	G^2	X^2	G^2	X^2	G^2	X^2	G^2	X^2
index	15.471	16.408	23.100	24.122	18.458	19.294	23.602	24.368
1	0.2450	0.2471	0.2722	0.2774	0.2703	0.2755	0.2746	0.2788
2	0.1959	0.1981	0.2257	0.2312	0.2210	0.2265	0.2268	0.2316
3	0.0347	0.0334	0	0	0	0	0	0
4	0.0832	0.0859	0.0802	0.0827	0.0751	0.0759	0.0841	0.0854
5	0	0	0.0062	0.0042	0	0	0	0
6	0	0	0	0	0.0199	0.0189	0	0
7	0.0656	0.0688	0.0649	0.0674	0.0562	0.0570	0.0674	0.0691

Asymptotically (t constant, $c \rightarrow \infty$), and when the data comes from the model (i.e. under the null hypothesis), the statistic -2lnLR or G^2 is distributed as a chi-square (χ^2) variable (e.g. Stuart and Ord p 1160). The degrees of freedom (d.f.) of this variable are equal to the degrees of freedom in the data (2^{t-1}-1 for 2 states, or 4^{t-1}-1 for 4 states) minus the number of statistically efficiently estimated parameters in the model (we discuss the distribution of "residual" statistics in detail in chapter 6). The asymptotic exact degrees of freedom will vary with the form of estimator used (see Stuart and Ord 1990, pages 1166-1172) and we will consider this further, along with the problem of sparseness (caused by finite data) in chapter 6. For example when fitting the 2-state Poisson model to the transversional changes of our 4 taxon rRNA data (as we have done in table 5.5), then the d.f. the G^2 statistic is 2^{t-1}-1 (= 7) - the number of optimised edge weights (5 for a binary tree, 4 for the star tree) minus any other fitted parameter (none in this case), leaving 2 d.f. for the binary, and 3 d.f. for the star tree. The X^2 statistic has the same asymptotic distribution under the null model, but weights deviations from a perfect fit in a different way, so the two statistics are unlikely to be exactly the same with sampled data (yet often agree closely). The other problem of course is tree selection, and not knowing the true tree (or model). Picking the tree model with the minimum observed fit will also effectively result in some, unknown, loss of degrees of freedom (e.g. Reeves 1992, Goldman 1993a).

When using either of these fit statistics, the adequacy of the assumption that these statistics are close to chi-square distributed with the specified degrees of freedom is data dependent. This assumption can be quite incorrect when the data is sparse, as can readily occur with sequences (chapter 4, discussion section). We consider this factor further in the next chapter when we wish to use this statistic to test quantitatively the hypothesis that the data fits the model. This factor does not invalidate the minimum G^2 statistic as a tree selection criterion, it is still asymptotically a statistically efficient optimality criterion (that is it best reduces sampling error when choosing the true tree). This statistical efficiency also extends to having the minimal standard errors on estimated parameters like edge lengths. Unfortunately, there is no clear indication as yet of when a ML estimator based on sequences will outperform other estimators which do not posses this property (e.g. OLS, or the minimum sum of OLS edge lengths criterion). It is worth noting that

under the model, other BAN (best asymptotically normal) estimators such as minimum X^2, will be equally efficient, and perhaps better with less data (Stuart and Ord 1990, p. 1162, Read and Cressie 1988). This does not directly tell us anything about the important properties of statistical efficiency with real sequences, nor about robustness.

We now discuss the relationship of the likelihood ratio statistic, or G^2, to the raw log likelihood returned by maximum likelihood tree inference programs such as DNAML and DNAMLK of Phylip 3.5 (Felsenstein 1993). The natural logarithm of the likelihood of the data given the model is,

$$\ln L_T = \sum_{i=0}^{2^{t-1}-1} \ln([s_i(T)]^{\hat{f}_i}) \qquad (5.3.1\text{-}2)$$

$$= \sum_{i=0}^{2^{t-1}-1} \hat{f}_i \ln(s_i(T)),$$

where as before $\hat{f}_i$ is the observed frequency of the i-th sequence pattern, and $s_i(T)$ is the probability of this pattern under the model.

Under many models it is necessary to calculate this likelihood for each pattern separately by summing the probability of observing such a pattern by all possible "pathways" of substitutions (i.e. summing over all possible assignments of character states to the internal nodes, Felsenstein 1981, Barry and Hartigan 1987a). If the mechanism of change is time reversible (Felsenstein 1981a, Barry and Hartigan 1987a) then these calculations can be made without needing to allocate a real root to the tree (Felsenstein 1981a, Barry and Hartigan 1987a)(an arbitrary root is often used for computational convenience). Hendy's theorem (Hendy 1989) shows that this is just what the Hadamard conjugation does for the time reversible models already discussed in chapter 2. The quantity lnL is always a negative number (since the probability is less than 1 of observing any sample from a multinomial), and for a given set of data, the more sequence sites and / or taxa that are added the smaller likelihood of the observed data (so the more negative lnL becomes). This last effect means lnL is not useful for comparing the likelihood of one model against another. Accordingly, Felsenstein in Phylip 3.5 (Felsenstein 1993) refers to this value as the computational likelihood. Maximising lnL is fine for optimisation on a specific data set, and gives the exactly the same parameter values as minimising the G^2 statistic.

We now look in detail at the likelihood ratio statistic or G^2. We define LR is the ratio of the maximum possible likelihood of the data (which is calculated assuming just the independence of sites) (L_U), to the maximum likelihood of the date under the model of interest (L_T)(others such as Stuart and Ord 1990 often cite this ratio the other way around. However, since it is almost always quoted as a logarithm, the only difference is one of sign). For the multinomial distribution, the maximum possible likelihood under any model (the unconstrained model) is,

$$L_U = \prod_{i=0}^{2^{t-1}} \left(\frac{\hat{f}_i}{c} \right)^{\hat{f}_i}, \qquad (5.3.1\text{-}3)$$

that is the likelihood under the unconstrained model is just the ML estimator of the probability of a site pattern under the multinomial model (s, see 4.2.2) multiplied with the probability of all other sites to give an overall probability of the data. So

$$\ln(L_U) = \sum_{i=0}^{2^{t-1}} \hat{f}_i \ln\left(\frac{\hat{f}_i}{c}\right), = \sum_{i=0}^{2^{t-1}} \hat{f} \ln(\hat{s}_i) \qquad (5.3.1\text{-}4)$$

Thus our likelihood ratio statistic (LR), which is most typically expressed as the natural logarithm of this ratio (lnLR), is just equation 5.3.1-4 minus equation 5.3.1.1 so

$$\ln LR = \sum_{i=0}^{2^{t-1}} \hat{f}_i \ln(\hat{s}_i) - \sum_{i=0}^{2^{t-1}-1} \hat{f}_i \ln(s_i(T)) \qquad (5.2.1\text{-}5)$$

$$= \sum_{i=0}^{2^{t-1}} \hat{f}_i [\ln(\hat{s}_i) - \ln(s_i(T))]$$

$$= \sum_{i=0}^{2^{t-1}} \hat{f}_i \ln\left(\frac{\hat{s}_i}{s_i(T)}\right), \qquad \text{recalling } f_i = cs_i$$

$$= \sum_{i=0}^{2^{t-1}-1} \hat{f}_i \ln\left(\frac{\hat{f}_i}{f_i(T)}\right) = \ln LR,$$

the number which we will routinely quote when working with the likelihood of evolutionary models (and 2 times this number is the G^2 statistic). Since the likelihood of the unconstrained model must always be equal to or greater than that of any other model which assumes independence of sites, the lnLR is always ≥ 0. Navidi *et al.* (1991) and Goldman (1993a) give similar expositions of the relationships between likelihood statistics.

An important point is that any specific model based on a multinomial sample (including any tree) can always be considered a submodel of the general, unconstrained model. Accordingly it is valid to compare the likelihood (or any other appropriate statistic) of any specific model against that of the general, unconstrained multinomial model. Thus if we know the distribution of the lnLR statistic, we can make a test of the adequacy of any sub-model (including any or all trees), and this in turn allows us to define things such as a confidence set of trees. The same point (i.e. all trees being sub-models of a more general model) has been made by Ritland and Clegg (1987), Bulmer (1991a), Navidi *et al.* (1991) and Goldman (1993a). Bulmer (1991a) made it in the case of the GLS SS statistic applied to pairwise distance matrices, while the other mentioned authors made it with specific reference to maximum likelihood tree selection. We will develop this theme further in chapter 6, where we consider tests of models of evolution.

Table 5.5 shows the maximum likelihood fit of our example data to the i.r. 2-state equi-frequency tree model of evolution (with all 4 possible tress evaluated). Notice how the external edges tend to be slightly longer (on average > 10% longer) than the estimates obtained by GLS (non-iterated) optimisation of fit to $\hat{\gamma}$. In contrast, with the ML criteria applied to sequences, the

internal edges are shorter than under GLS by an average of over 30%. As with GLS the resulting statistic strongly suggests that the data was not generated under the assumed model, cautioning us in our interpretation of accepting any tree as unambiguously better supported than any other. Application of the method of maximum likelihood and the i.r. 2-state equi-frequency model to similar sequences using Lake's (1988) alignment sees an alternative tree chosen as optimal, in this case the so called eocyte tree.

In chapter 6 we discuss the power divergence family of goodness-of-fit statistics that includes not only G^2 and X^2 (the Pearson statistic) but also other less well known statistics, and a continuum of intermediates. No one of these statistics is clearly always the best at detecting deviations of sampled data from expected data, and we expect a similar trend may occur in the robustness of tree selection. For example, with many of the classically considered situations the X^2 has better discriminatory power to local alternatives than G^2, and also provides a better approximation to the asymptotic chi-squared distribution under the null model for sparse data (Read and Cressie 1988). In table 5.5 we have optimised the tree models according to X^2 statistic. The solutions are very similar to those obtained by likelihood (i.e. G^2); X^2 appears to find the data slightly less well fitting than G^2; the X^2 optima have made all the external edges slightly longer than in the case of G^2, and have shrunk the internal edges by varying amounts. These explicit optimal fits for X^2 were about 0.5 lower than if one took the X^2 value at the G^2 optima (and vice-versa for the G^2 statistic on X^2 optima). A study of the optimal $\mathbf{s}(T)$ for different statistics and different trees showed that X^2 was particularly sensitive to the deviations of the "parsimony patterns" not in the tree and would increase the fit of these at the expense mostly of the constant sites. Consequently, $s(T)_0$, optimised by the X^2 criteria was 6 out of 1352 smaller than predicted by G^2.

A great advantage that likelihood has over X^2, and most other measures of goodness-of-fit, is that likelihood statistics (including G^2) require only the calculation of the expected probability of site patterns which are observed in the actual data. Since the number of possible site patterns increases exponentially with t (the number of taxa), X^2 and many other goodness-of-fit measures are unlikely to be calculated exactly for more than 15 taxa (with 4 states, or 30 taxa with 2 states). This does not mean they are without use if they have desirable properties, since many of the most interesting and hard to resolve phylogenetic problems are well posed with 10 or fewer sequences (especially if we replace whole clades with adequately determined ancestral sequences). Hadamard conjugations are well suited to using these other goodness-of-fit statistics, since they do calculate all pattern probabilities, at a speed which cannot be matched by any other method of calculation.

It is of course possible to replace the i.r. Hadamard conjugation's with the rates across sites conjugations and accordingly make ML estimates under a model with a specified, stationary distribution of rates across sites. Such calculations are discussed and illustrated in the section after next.

5.3.2 Finding the maximum likelihood point of a specific tree

Under i.i.d. Markov models of sequence evolution, there is no closed form solution to finding the maximum likelihood point (i.e. the set of optimised model parameters) for any tree. The result in the previous section shows that as $c \rightarrow \infty$ with t fixed, then iterated generalised least squares applied to $\hat{\gamma}$ converges to the maximum likelihood solution; that is we can solve for the optimal point at any given iteration by finding the minima of GLS SS using the previous iterations covariance matrix, but we then have to update the covariance matrix and in doing this the solution will shift, but should converge. Similar procedures exists for finding optima, even when sequences are not long. Probably the best known is the Newton method for finding the minima of a function, which is the solution to a series of non-linear equations, $\frac{\partial f}{\partial x_i}(x) = 0, (i = 1,..,n)$(e.g. see Minoux 1986, p. 94). The method uses first and second order partial derivatives of the function of interest. In the case of finding the $\gamma(T)$ vector which minimises the lnL statistic, in each case make the next iteration,

$$\gamma(T)_{itt+1} = \gamma(T)_{itt} - \frac{f'[\gamma(T)_{itt}]}{f''[\gamma(T)_{itt}]}, \qquad (5.3.2\text{-}1)$$

$$= \gamma(T)_{itt} - f'[\gamma(T)_{itt}]^{-1} f'[\gamma(T)_{itt}],$$

where itt stands for the iteration just completed, $f'[\gamma(T)]$ is the vector of the partial derivatives of $\gamma(T)$ with respect to the log likelihood function, and $f''[\gamma(T)]$ is the matrix of second order derivatives of $\gamma(T)_i$ and $\gamma(T)_j$ with respect to the likelihood function. As with previous sections we are only interested in the entries in $\gamma(T)$ which may correspond to an edge in a tree, so that we exclude γ_0. The matrix $f''[\gamma(T)]$ of second order partial derivatives is known as the Hessian matrix in numerical mathematics. In statistical literature when $-f''[\gamma(T)]$ is the matrix of negative the expected values of the second order partial derivatives of the likelihood function, it is called the information matrix (Agresti 1990, p 114). Its relationship to the variance-covariance matrix of the estimated weighted tree, $\gamma(T)$, is $\mathbf{V}[\gamma(T)] = 2(f''[\gamma(T)]^{-1})$ (Stuart and Ord 1990, p. 638). With finite samples the lnLR statistic does not behave as predicted asymptotically, and these confidence intervals should be treated as approximate and can be either too large or too small (more on this in chapter 6).

In certain situations we can obtain a tractable and analytic expression for both f'[γ(T)] and f''[γ(T)]. This turns out to be the case with the identical rates Hadamard conjugation models discussed in chapter 2. The vector of first derivatives of γ(T) is,

$$\begin{aligned} &= \frac{\partial \ln LR(\mathbf{s})}{\partial \gamma_j} \\ \mathrm{f}'[\gamma(\mathrm{T})] &= \sum_{i=0}^{2^{t-1}-1} \frac{\partial f(s_i)}{\partial \gamma_j} \\ &= \sum_{i=0}^{2^{t-1}-1} -\frac{\hat{s}_i}{s_i(T)} \times s_{i\oplus j}(T), \end{aligned} \tag{5.3.2-2}$$

where $f(s_i)$ refers to the per cell contribution to the lnLR statistic, that is $\hat{s}_i \ln[\hat{s}_i/s_i(T)]$ (eq 5.2.1-5). The ⊕ symbol refers to the adding of indices in modulo 2, also called the exclusive 'or' (i.e. add the indices together without carry over to the next place to the left). For example with 2 colours and 4 taxa s_3 has binary index 011, while s_5 has index 101, so $s_{3\oplus 5}(T)$ has index 011⊕ 101 = 110 = 6, i.e. s_6. The Hessian matrix of γ(T) has entries,

$$\begin{aligned} &= \frac{\partial^2 \ln L(s)}{\partial \gamma_j \partial \gamma_k} \\ \mathrm{f}''[\gamma(\mathrm{T})] &= \sum_{i=0}^{2^{t-1}-1} \frac{\partial^2 f(s_i)}{\partial \gamma_j \partial \gamma_k} \\ &= \sum_{i=0}^{2^{t-1}-1} -\frac{\hat{s}_i}{s_i(T)} \times s_q(T) + \frac{\hat{s}_i}{[s_i(T)]^2} \times [s_p(T)]^2, \end{aligned} \tag{5.3.2-3}$$

where $f(s_i)$ refers to the per cell contribution to the lnLR statistic, that is $\hat{s}_i \ln[\hat{s}_i/s_i(T)]$ (eq 5.2.1-5), and where p = i⊕j, and q = i⊕j⊕k (i.e. the same order insensitive adding of binary indices without carrying over used above). These results were developed with Dr Mike Hendy and proofs will appear elsewhere. A well known result in numerical analysis is if the Hessian matrix is always positive definite, then there will be but one optima within a model. And we are hoping to use this property to explore the conditions under which the Hessian matrix is not positve definite, and so allows for multiple likelihood optima per tree as shown by Steel (1994).

Another use of the Hessian is that it asymptotically gives the information matrix of Fisher (Stuart and Ord 1990, p. 638), which in turn asymptotically yields the variance-covariance matrix (with a tendency to underestimate its entries, since it only takes account of terms of the order 1/n). It will be useful to be able to calculate this covariance matrix, as it can then be compared with $\mathbf{V}[\hat{\gamma}]$ and we may understand more of how ML on sequences becomes more efficient than other tree selection methods (including those applied to s(T)). Entries from this matrix are useful for making statistical tests, and Felsenstein (1993) gives an example with a univariate normal test of the hypothesis that an edge in the tree is greater than zero in length. Other tests described in chapter 6 make particular use of not only the variances, but also the covariances.

In March to July 1993 I ran many ML analyses of Horai *et al.*'s (1992) 5kb of mtDNA sequence. Some of these results are presented later, while others appear in Waddell and Penny (accepted 1993, to appear 1996). One interesting general result was that the extra time required to run models with up to three independent edges per tree was not prohibitive, while the addition of a parameter to allow for a distribution of rates across sites, could even speed up the rate of convergence to the optimum (we started form multiple edge lengths and were confident only one optima existed when rates across sites were considered). There is no single test which will accurately reveal the relationship between number of parameters and time to find the overall optimum, since this will be data and model dependent (its arbitrary, for example, how long the next external edge added to a tree will be, but this edge's length can influence the time to find an optimum). What we observed was that for models with up to 6 taxa (hence up to $9 \times 3 = 27$ plus distribution of rates across sites parameters, all of which are independently optimised) the time to convergence was approximately linear with the number of parameters. As such it should usually not be greatly more expensive to optimise parameters such as tr / tv ratio, or the shape of a distribution of rates across sites, than it is to add in extra edges to a tree (i.e. add in more taxa) assuming a Newton type method is being used.

In practice we used a quasi-Newton method to find optima according to likelihood. A quasi-Newton method often has very similar performance to Newton methods, but it calculates the Hessian matrix heuristically (i.e. does not have a closed form expression for it) (see Minoux 1986). (Note our usage of the term Newton method implies all parameters are optimised jointly, as opposed to optimising one parameter at a time with a Newton-Raphson type procedure). We also used a conjugate gradient method to check it would find the same optima found with the quasi-Newton method. In general the quasi-Newton method was approximately 2 to 4 times faster than the conjugate gradients method. The biggest problems encountered with both methods were not so much divergence, but stopping short of the ML point when the likelihood surface was flat, or else edges were long and short, and there were strong correlations (positive and negative) between edge lengths with respect to the change in likelihood (this is expected given the results in chapter 4, which show that entries in $\mathbf{V}[\hat{\gamma}]$ can have strong correlations). We found this type of problem (termination of search before minima was found) was apparently much more common in real data under models without a distribution of rates across capacity. We always performed multiple runs to check that each tree's ML point was found by different methods from different starting points (parameter values). The specific numerical solution implementation used is the solver package that comes with the program Microsoft Excel version 4 for Windows.

5.3.3 Branch and bound of maximum likelihood

Branch and bound is a search procedure which guarantees to implicitly find the optimal tree by methodically eliminating trees which could not be as short as the shortest tree found to date in a search, and only explicitly evaluating those trees which remain. One form of it, called the singles algorithm, was first implemented in phylogenetics by Hendy and Penny (1982). Other distinct variants such as 'Turbo tree' (Penny and Hendy 1987) or 'pairs algorithm' (Penny and Penny 1988) are useful depending upon the data structure. Since many ML analyses are run on 15 or less taxa, this method could potentially be very useful if it could be implemented effectively with likelihood. Here we consider some of the issues involved, and put forward a conjecture, which if correct should make branch and bound much more useful in this context. (see chapter 1 for an introduction to branch and bound in phylogenetics, while Swofford and Olsen (1990) gives a useful illustrative example).

Adding an extra taxon cannot under any circumstance decrease the minimum lnL statistic of a tree under a specific model. This can be seen clearly if we look at the likelihood site by site. As soon as another sequence is added we are effectively adding a external edge to the tree (and in the process bisecting an existing edge). The likelihood of a site cannot increase, since there is at least one more state per site, making the chance of the model generating exactly that pattern less likely. The least the likelihood can decrease is when adding a sequence identical to one of the existing taxa. In this case the likelihood will change by zero. This makes the method amenable to branch and bound, with the likelihood of any subtree of tree(X) being an upper bound on the likelihood of tree(X), (and similarly for the lnL statistic).

A simple 'singles' branch and bound algorithm is to:

(a) Find a near optimal tree for *t* taxa by a reliable heuristic method. This provides our initial upper bound on the likelihood.

(b) Start the branch and bound process by selecting 4 taxa. Measure the likelihood of each of the three binary trees for these 4 taxa. If any of these trees has a lower likelihood than the bound found in (a) then all trees that have this tree as a subtree are excluded from having a likelihood as high as the optimal *t* taxon tree. We do not need to explicitly evaluate the likelihood of each of these trees in order to eliminate them as candidates for the optimal tree for t taxa.

(c) Select the next taxon to add, and expand each of the remaining 4-taxon trees out into 5-taxon binary trees by adding this taxon into each edge. Find each trees likelihood, and excluded any which has a likelihood lower than the heuristic upper bound.

Continue this process of adding taxa, and checking all remaining trees, until all possible trees have either been explicitly checked, or eliminated because a subtree had a lower likelihood than the bound. If, at any time, a t taxon tree is found to have a lower likelihood than the current bound, replace the bound with this tree and continue the search. (In the preceding sections, finding a trees likelihood means finding the maximum likelihood the data could have for this tree, given the specified mechanism of change).

Let us now look at the application of a branch and bound approach to the data set of Horai *et al.* (1992), in order to learn more about its usefulness. Our edit of this data set is minus just the small proportion of sites in which one or more taxa show an insertion or deletion. To begin with we will use the 5 parameter i.i.d. model of Felsenstein (1993) of the program. This program is used because it is currently the most widely used ML tree selection method in phylogenetics. We searched for the tree with the highest likelihood of producing the data using the program's heuristic options of placing each taxon at the best location in the growing tree (with taxon addition order generated at random), followed by the global rearrangement option. This found the tree ((((chimp, pygmy chimp), human), gorilla), orangutan, siamang) as expected, and this tree had lnL value -14972.6.

A crucial decision to be made in a branch and bound search is which 4-taxon subtree to start with? Branch and bound in the parsimony program PAUP starts by finding the 3-taxon tree that has the greatest length. This seems a reasonable starting point as one would expect that the longest 4 taxon tree will probably have these three species in it, plus another species (which would be expected is either closely allied with the tip of the longest external edge on the 3-taxon tree, or else has a long external edge itself, or some combination of these two features).

For parsimony searches, a maximally long 4-taxon tree is the best candidate to most immediately start eliminating subtrees as longer than the heuristic bound. The same goes with likelihood, except here we would look for the 4-taxon tree most unlikely to have generated the data in order to most immediately aim to eliminate non-optimal trees. The only way guaranteed to find the 4-taxon tree which is most unlikely is to find the maximum likelihood point of all possible 4-taxon trees. There being 3 times $\binom{4}{t}$ (t choose 4) of these, then this can incur an appreciable expense as t becomes large. In our example we check all 45 four taxon trees, and find that the one with the lowest lnL value is the tree ((siamang, gorilla), orangutan, human) (lnL = -13,225.1). This tree also has the largest sum of edge weights. Indeed there is a high correlation between likelihood and sum of edge weights (figure 5.2). Further, for the three binary trees from the same subset of four taxa this relationship always held (i.e. rank of lnL was always opposite that of sum of edge weights). This relationship suggests that the most unlikely four taxon tree will be a rearrangement of the four taxon subtree of the true tree having the greatest sum of edge lengths. Figure 5.2 also shows the wide range of likelihoods spanned by the various 4-taxon subtrees (from -10,428.1 to -13,225.1, nearly 1,800 lnL units).

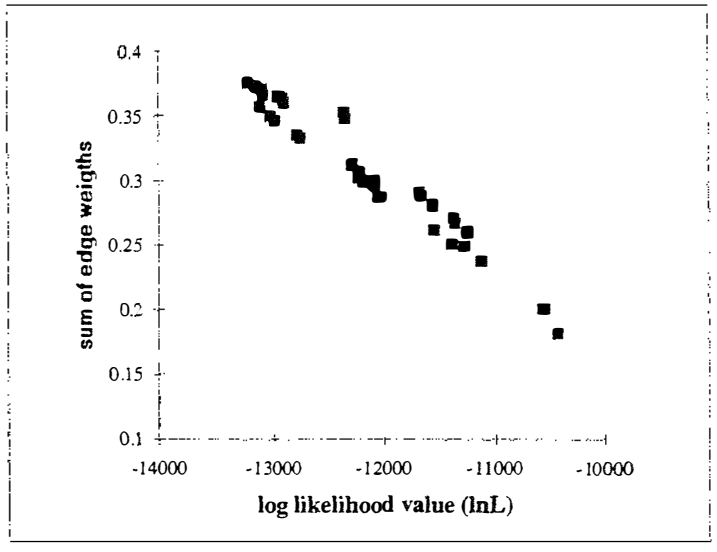


FIGURE 5.2. A plot of likelihood versus sum of edge lengths for all possible 4-taxon trees from the set of six hominoid species mtDNA sequences in Horai *et al.* (1992). Notice the high correlation, but not quite perfect rank (except within each set of four taxa). The tree with the lowest likelihood was ((siamang, gorilla), orangutan, human) (lnL = -13225.1), while that with the highest was ((pygmy chimp, common chimp), human, gorilla) (lnL = -10428.1). (This correlation is also interesting because it shows that the minimum-evolution or OLS distance tree with minimum sum of edge lengths, Kidd *et al.* 1972, is a criterion which has a probable analog in ML, i.e. a criterion of don't pick the best fitting tree, but rather that with the minimum sum of edge lengths).

In this example none of the possible 4-taxon trees has a worse likelihood than our bound on the best six taxa tree. Expanding the 4-taxon tree with the lowest likelihood out into the possible 5-taxon trees by adding the remaining taxon having the longest external edge (chimp), gives five trees. Of these, the one with the worst lnL value of -14,562.1 was the tree ((chimp, siamang), gorilla, (orangutan, human)). Of all the possible five taxon trees with the set of taxa (chimp, human, gorilla, orangutan, siamang) (that is missing just pygmy chimp) the best was the tree (((chimp, human), gorilla), orangutan, siamang) with lnL -14,346.5.

Unfortunately, this search has been unable to cut off any avenues implicitly. More importantly, however, the major cause for this becomes apparent. While the lnL difference between the best and the worst five taxon trees for the 5-taxon set considered is a considerable -14,346.5 – -14,562.1 = 215.6 lnL units, the cost of adding even the shortest edge in the 6 taxon tree (the edge linking pygmy chimp into the tree) is a substantial -14,346.5 – -14,972.6 = 626.1 lnL units. This may be thought of as approximately the probability of evolving the one sixth of the data consisting of the pygmy chimp sequence of ≈ 5 kb.

The key issue appears to be able to put a reasonable bound on the minimum amount the overall likelihood will decrease when adding an extra taxa. In the case of one successful branch and bound parsimony analysis program (e.g. PAUP, Swofford 1993), the cost of the singleton changes (states that appear in just one taxon) are readily calculated and usually substantially tighten the bound. Further, by careful bookkeeping of the compatibilities between characters, a further increment of the bound can be made upon specific subtrees. With likelihood we would

desire tight bounds for both these factors. However making this aim difficult is the fact that unlike parsimony, with likelihood columns of constant state also have a "cost" of evolving on the tree which is a function of the sum of tree edge lengths (the likelihood of the constant sites will only remain static if the length of the external edge adding in the new taxon is zero).

For a singleton state (i.e. the new taxon has a state at that site which no other taxa exhibits) not already in the tree we can estimate a loose lower bound for how much it must lower the lnL statistic to add it in under a stationary time reversible model. The least it could lower the log likelihood would be if the external edge for the taxa it belonged to was going to be given infinite length, so that its probability of changing to the new state was just $1/\pi_i$, where π_i is the expected frequency (under the model) of the state it must change to. Thus the lnL must increase by $\sum_{i=1}^{4} f_i \ln(1/\pi_i)$, where f_i is the frequency of the *i*-th singleton pattern. This may be a very loose bound, and unless the external edges to be added are fairly long, might make very little difference (unfortunately being a continuous variable influenced by all the data, it is difficult at present to see any way of putting any true bound on the exterior edge length for a sequence yet to be added to the analysis). For example, table 1.1 lists the base composition of these sequences. Accordingly our estimated loose bounds for adding in these sequences are:

H: -141.48 lnL units, C: -49.54, P: - 35.32, G: - 174.45, O: -333.59, and S: -351.24. Clearly the bound for adding in the pygmy chimp sequence (-35.32) falls well short of the -626.1 required to implicitly exclude certain trees. It is not difficult to envisage that branch and bound on likelihood can work, the problem is getting it to work optimally and usefully (for example if we take the data set of 16S-like rRNA data from chapter 3, find an optimal tree under likelihood, then remove mouse and rice, specify a very non-optimal 26 taxon tree and measure its likelihood, it will exceed that of the best tree).

By a similar argument, we should also be able to calculate the minimum change in likelihood for any character which implies at least one more parsimony change on the tree. The likelihood of a particular site evolving on a tree is the sum of the probabilities of all the different ways a site could evolve on a tree (i.e. summed over all possible internal node state assignments). We need to show that this quantity cannot decrease by less than 1 upon the frequency of the state about to be added (or more conservatively the most frequent state to take into account alternative interior node assignments); we do not see a straight forward proof at present.

An alternative approach is our conjecture that the lnLR statistic may also act as a bound, and if so, the difference between subtrees and full *t* taxa trees by this measure is usually much less, making it a useful proposition for branch and bound (if it is indeed a bound). We had considered this to be promising earlier, especially since the lnLR statistic for the ((((chimp, pygmy chimp), human), gorilla), orangutan, siamang) tree was typically lower than that of the majority of four taxon trees for a variety of mechanisms of evolution with likelihoods calculated via Hadamard conjugations. While the six taxon tree with the lowest lnLR is indeed also the maximum likelihood tree, it does not hold that the lnLR of a subtree must be higher than that of the whole tree. Consider that the lnLR statistic has two parts $\ln L_u$ - $\ln L_T$ (where T is the tree model

and u is the unconstrained multinomial model), and ignoring the combinatorial term which is identical in each case, then $\ln L_u$ can be expressed $\sum_{i=0}^{x} \hat{f}_i \ln(\hat{f}_i)$ - cln(c) (where c is the sequence length). The term cln(c) is not a problem because it is constant for any number of taxa as long as they all have the same sequence length. However the summation term decreases as the number of taxa grows (its decrease is analogous to the decrease in the significance of a X^2 test as the amount of data remains the same but is broken down into more subclasses). What we need to show is that this term cannot decrease more quickly than the drop in log likelihood that must be incurred by adding the remaining sequences to any tree model of evolution.

Consider again the unconstrained model, the overall maximum likelihood model with a parameter for every pattern encountered. For *t* sequences this models likelihood of this data is equal to the number of ways of choosing the different patterns observed multiplied by the probability of observing anyone of these sequences under the unconstrained model. That is $\ln L_u = \frac{c!}{\prod_{i=1}^{x} f_i!} \prod_{i=1}^{x} \left(\frac{f_i}{c}\right)^{f_i}$, where f_i is the frequency of the *i*th of x patterns, and c is the total sequence length, and N is the number of distinct patterns observed in the data. We will call the logarithm of this quantity $\ln L_{un}$. It is a well known result that $\ln Lu_{n-y}$, where -y is removing some of the sequences, is always less than or equal to $\ln Lu_n$. The unconstrained model adds in extra parameters at will to explain extra data, yet the tree model (with log likelihood $\ln L_T$) is strictly a submodel and is constrained to adding in at most a linearly increasing number of parameters of a fixed type and relation to one another. Consequently it seems reasonable to hypothesise that, $\ln Lu_n - \ln Lu_{n-y} \geq \ln Lt_n - \ln Lt_{n-y}$ would hold, so that the difference $\ln Lu_n - \ln Lu_{n-y}$ can act as a bound on the cost of adding the remaining y sequences. Rearranging this inequality yields $\ln Lt_{n-y} - \ln Lu_{n-y} \geq \ln Lt_n - \ln Lu_n$, or in other words $\ln LR_{n-y} \geq \ln LR_n$, (so $G^2_{n-y} \leq G^2_n$) which brings us back to our previous speculation. At present we know of no general proof that this condition must hold, but as yet we have not seen it violated with real data, nor have we produced a counter example with model data.

To evaluate the fit of G^2 we will use the extended generalised Kimura 3P model allowing the rate of transitions and type 1 and 2 transversions to vary independently on all edges of the tree (see chapter 2). In addition the distribution of rates across sites is modeled as a Γ distribution (with all parameters in the model, including shape parameter *k*, simultaneously optimised using quasi-Newton methods, and then checked for convergence with a conjugate gradient method). (These results are part of those prepared for Waddell and Penny 1995, and we use them as the program DNAML does not offer the G^2 statistic or optimise a distribution of rates across sites.) Table 5.6 shows the results of the G^2 statistic optimised on all possible five taxa trees for the sequences of human (H), common chimp (C), gorilla (G), orangutan (O) and siamang (S) from Horai *et al.* (1992). The best fit obtained for the full 6 taxa data set with these sequences was G^2 = 303.20, with k = 0.351.

Table 5.6 Fit of five taxon trees evaluated under the generalised Kimura 3P model with a Γ distribution of rates across sites (shape parameter, *k*).

Tree	HC)G)O	CG)H)O	HG)C)S	HC)S)G	HC)O)G	HS)C)G	GO)H)C	CO)H)G
k optimised	0.35	0.31	0.31	0.28	0.27	0.25	0.25	0.24
G^2 statistic	171.3	214.2	215.7	252.5	258.0	288.2	288.2	304.0
Tree	HO)C)G	HO)S)C	HG)S)C	HS)G)C	HS)O)C	HO)G)C	HG)O)C	Tstar
k optimised	0.24	0.23	0.23	0.23	0.23	0.23	0.23	0.23
G^2 statistic	304.7	309.3	309.6	309.6	309.8	310.0	310.9	310.9

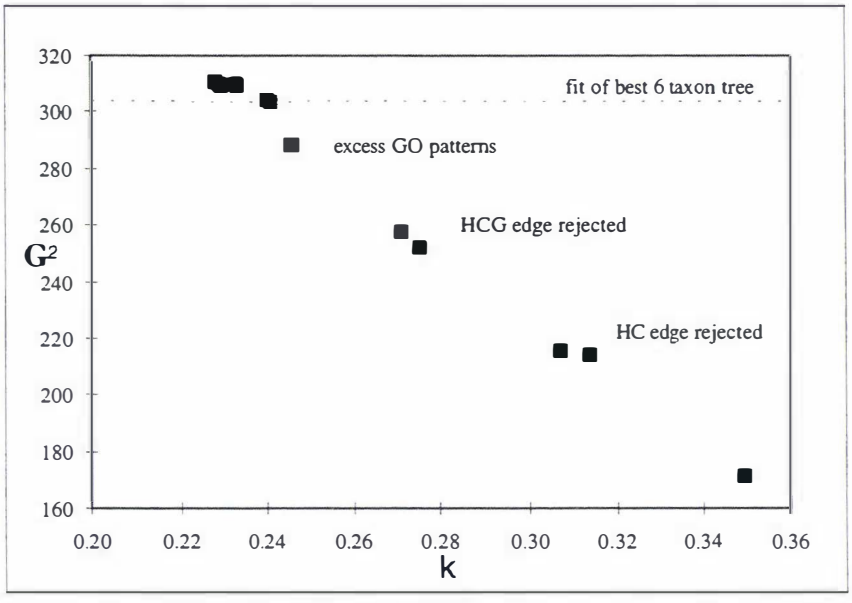


FIGURE 5.3. A plot of the G^2 (=2lnLR) statistic vs sum of edge lengths for all possible 5 taxon trees for the set of sequences human (H), chimp (C), gorilla (G), orangutan (O) and siamang from Horai *et al.* (1992). The model used to evaluate these data was the extended generalised Kimura 3P model, allowing a gamma distribution of rates across sites. Notice that the G^2 statistic for more than half the possible 5 taxon trees are worse than the 6 taxon tree with the best G^2 statistic. Notice also the strong negative correlation between G^2 and k (the shape parameter of the distribution of rates across sites).

Plotting the data of table 5.6, as in figure 5.3, better shows interesting features. Nine of the binary 5-taxon trees exceed the G^2 statistic of the best 6-taxon tree, and if our hypothesis that G^2 is a bound on the best fitting tree is correct, then we do not need to further evaluate expansions of these trees (to know that the optimal tree is not one of them). The cost of not having the grouping (HC) in the tree appears to be approximately 43 G^2 units, while not having the group (HCG) is about 82 G^2 units. Together with the fit of the best five taxon tree, these do not quite add up to the fit of the worst 5 taxon tree, the star tree, suggesting that the cost of rejecting these edges in a tree is balanced by better explanation of some parallelisms and convergences. Indeed a large number of parallelisms and convergences between the gorilla and the orangutan sequences explain why trees with this grouping are significantly better fitting than expected (i.e. the fit of the star tree) by 22.7 G^2 units. In turn this feature of the data suggests that the model may not be fitting within expectations (more on this in chapter 6). The other striking feature of this data is the strong negative correlation between *k* and G^2. An approximate 95% confidence interval about

k for the 5-taxon tree is 0.24-0.46 (from a likelihood ratio test), with the value of k from the star tree falling outside this interval. These two features together argue that it is not wise to estimate a parameter of the data such as k on the star tree. This finding contradicts a suggestion made by Yang *et al.* (1994), who suggested using the star tree to evaluate ancillary parameters of interest to molecular biologists. We suggest instead that any additional parameters such as transition to transversion ratios, distribution of rates across sites, should always be evaluated upon the best fitting tree found. And of course, take precautions in estimating ancillary parameters when there is good evidence that the method being used may be selecting an incorrect tree.

Next, consider the question of whether likelihood must decrease as more taxa are added yet there is simultaneous reoptimisation of all model parameters each time more data is added (this must hold if likelihood is to be a bound). The answer is yes, the likelihood cannot increase. Suppose that the transition to transversion ratio is fixed and not updated each time more data is added. This situation is identical to that discussed earlier where the new sequences must decrease the overall likelihood. To illustrate that the likelihood must decrease if there is reoptimisation with any new data being added, consider that with four sequences the tr / tv ratio has been optimised for a specific tree model. As more sequences are added, but initially holding the tr / tv ratio constant, then the likelihood must decrease. Let us assume that the added data fits best with a reduced tr / tv ratio. If the tr /tv ratios now reoptimised, the likelihood of the data will decrease. However, it cannot decrease to less than the likelihood of the data on the previous tree, since the likelihood of the data before the addition of the new sequence can at best be equal to its former likelihood (only if the tr / tv ratio has not changed). The same argument can be applied to any more complicated model (e.g. with a distribution of rates across sites, and / or optimisation of further parameters in the substitution matrices associated with each edge in the tree). A practical problem which could upset this relationship is if numerical methods failed to find the best optimum with a subset of taxa, but found a better optimum when more taxa was added (and of course this relationship is also invalid if the mechanism of evolution has altered).

Its also useful to consider if branch and bound on likelihood can be performed when there are sites with gaps or ambiguous states. If a gap at any site is treated as missing data (as they effectively are when the mechanism does not model insertion-deletion processes), then gaps do not change the likelihood of that site on the new tree with extra tips (Felsenstein 1981a). Thus they do not interfere with the tenant that the likelihood cannot increase as more data are added. When an ambiguous state appears at a site (e.g. code R, meaning the state is either an A or a G) and is left in the model, then the likelihood of that site is the sum of the probabilities of all the patterns consistent with that ambiguity. For example the likelihood associated with the site AAGR is the probability of pattern AAGA plus the probability of the pattern AAGG. If states consistent with R have probability of one occurring, then the likelihood remains unchanged (essentially like the N, –, or ? state), else the likelihood of this pattern must decrease relative to the pattern on the subtree AAG (as long as the whole new sequence is not identical to a previous one). It is possible to evaluate the probability under the unconstrained multinomial model of sequences with missing data (thus $G^2 = 2(\ln L_U - \ln L_T)$, thus if the G^2 statistic is a bound, this

does not present a great problem.. However, with missing data, G^2 will not have the simple form of the last line of 5.2.1-5, unless we designate sites with missing states as distinct patterns (which means more special summations when using Hadamard conjugations to estimate likelihoods).

A useful application of branch and bound searching is to find all trees within a certain likelihood of the optimal tree. One purpose of such a search is to construct a confidence set of trees, such that the overall probability of the true tree being amongst them reaches some level (see chapter 6). In such a search, it is usual to replace the lower bound on the maximum likelihood of the *t* taxa tree with this number minus some constant (asymptotically this constant can be estimated from a χ^2 distribution, to obtain a conservative confidence set). Such a search for a confidence set of trees can be also be done dynamically, coincident with the initial search for the best tree; i.e. update the bound for collecting suboptimal trees to whenever a *t* taxon tree with a higher likelihood is found (e.g. as Hendy and Penny 1982 did with parsimony). Doing this avoids collecting trees which will not be in the final confidence set.

The findings here suggest that until we can tighten our bounds (especially if it can be proven the G^2 statistic can act as a bound), then branch and bound of likelihood has the best chance of succeeding when the tree to be analysed has long internal edges and short exterior edges, with few parallel changes to lend support to edges not in the tree. In addition, a prediction worth evaluating is that branch and bound on likelihood might work best for data sets with more character states (e.g. amino acids) as they will have more singletons (for use with the loose bound) and fewer parallelisms in general. However, in order to make full use of branch and bound with likelihood it is clear there must be some breakthrough in our present understanding of the minimum amount that the character states in the unadded taxa must decrease the likelihood. I am presently looking to prove or disprove that G^2 is a bound, with Mike Steel and Jaxk Reeves help.

5.3.4 Maximum Likelihood with a distribution of rates across sites.

Here we give an overview of current methods of making likelihood calculations, taking into account unequal, but fixed relative rates of substitution. In the next section, models where the distribution of rates across sites is non-stationary are developed (with respect to the relative rates of sites). The most basic mechanism with unequal rates at different sites is the invariant sites type of model. In this model, some sites simply cannot change (they are said to be invariant), while all other sites are assumed to be i.r. and i.i.d. distributed (just like the original model of Felsenstein 1981a). Hasegawa *et al.* (1985) first implemented this model (although details in their 1985 paper are vague; Churchill *et al.* (1992) give a more detailed treatment). Under these models, the proportions of bases in the invariant sites is assumed to be equal to those in the variable sites, so there is just one parameter more to optimise in these models, namely p_{inv} (the proportion of invariant sites). Letting $\mathbf{s}_{all}$ be an ordered vector of the probability of each possible

sequence pattern under the model (an i.i.d. process of sequence evolution, a tree with edge lengths and a proportion of sites which cannot change), then $\mathbf{s}_{all} = (1-p_{inv})\mathbf{s}_{var} + p_{inv}\mathbf{s}_{inv}$. The first vector after the equals sign is just the probabilities under the tree model, times the proportion of variable sites, and the vector $\mathbf{s}_{inv}$, will be a vector of all zeros, except for those entries corresponding to the probabilities of all sites showing the same pattern, which take value π_i, where i is the frequency of the i-th state. The overall likelihood is the product of entry $s_{all}(j)^{f(j)}$ over all patterns j which are observed in the sequences being analysed (where s(j) is the probability of the j-th sequence pattern, and f(j) is the number of sites in the data having pattern j).

The log likelihood for the invariant sites model is just as simple, replacing the product with a sum, $\ln L = \Sigma\, f(j)\ln[s_{all}(j)]$, summed over all patterns j, observed in the real data. To find a likelihood optimum on each tree numerical methods are used (see earlier). Following this there is the usual search across trees. To maximise the probability of finding the tree model which is most likely to have generated the data under the assumptions made, in theory all parameters must be optimised for each tree evaluated. The invariant sites model is useful for a number of reasons. It mimics reality, because there are sites in functional molecules which cannot accept substitutions due to deleterious effects. Also, the invariant sites model, to a first order, seems to well approximate any other distribution of rates across sites. This can be expected after comparing distances estimated under an invariant sites model, when say a Γ distribution model generated the data (e.g. see figure 3.4). Since a plot of rates across sites must be normalised to one (see chapter 2), then about the two most different looking distributions you can have are one with 50% of sites having rate zero while the remainder have an identical rate (giving a U-shaped distribution), versus a continuous L shaped distribution like a Γ. Yet we have found that these two models well approximate each other in terms of the frequencies of sequence patterns one would expect to see, (for example a 50% invariant sites distribution approximates to a gamma distribution with shape parameter $k \approx 0.5$). Our own experience developing maximum likelihood with a Γ distribution of rates across sites (Waddell and Penny 1995, this chapter) is that these models tend to return generally similar likelihoods, compared to an i.r. model. One way of integrating these observations is through the Hadamard conjugation, where near additivity of distances translates to closely agreeing observed similarities on all pathsets (the r_i entries), which in turn translates to similar **s** vectors, which means nearly equivalent likelihoods. The best additivity of invariant sites distances to Γ distributed distances, often comes with larger absolute values of the invariant sites distances (see figures 3.4-3.6), and this helps explain why the edge lengths estimated by the invariant sites model tend to be larger than those estimated under a Γ model.

The logical step towards a more complicated model is to specify two distinct (and non-zero) rates of substitution. Proportion p_1 of sites evolve by an i.i.d. mechanism at rate $\lambda 1$, while proportion $(1-p_1)$ evolve at a second rate $\lambda 2$ (but by an identical mechanism). The vector of pattern probabilities here is just $\mathbf{s}_{all} = p_1\mathbf{s}_{\lambda 1} + (1-p_1)\mathbf{s}_{\lambda 2}$. Since the rate of substitution is fixed in its

ratio, then the weighted tree for the set of sites with rate $\lambda 1$, is exactly the same as the weighted tree used to generate the probability of site evolving at rate $\lambda 2$, except all its edges are bigger than those in the first tree by the factor $\lambda 2 / \lambda 1$. Under this model there are more parameters to optimise than under either the i.r. model, or the i.r. plus invariant sites model; we need to optimise the parameters p_1, $\lambda 1$, and $\lambda 2$. This approach can be extended for more and more rate classes. There are two disadvantages in continually adding more discrete rate categories: (1) more computational effort, and (2) as we add more parameters to our model, the variance of our most important parameters (e.g. edge lengths on the tree) increases. An advantage is that we don't need to make so many simplifying assumptions, and it may turn out to give very useful information on the distribution of rates across sites.

More parameters per tree model can also lead towards the likelihood of different tree models becoming more similar (and the more similar they become, the more difficult it is to reliably say this tree is better than that one). There is one commonly available program at present which allows one to estimate the likelihood under this type of model. It is DNAML (Felsenstein 1993), unfortunately in its present version the proportions and rate of each class of sites must be specified (it does not perform any optimisations of these parameters). This model also allows some correlation of rates between adjacent sites; to specify an i.i.d. model, this parameter should be set to zero. (To specify an invariant sites model, specify no correlation in rates of adjacent sites, and two rate classes, one with rate zero, and the other any positive relative rate).

A useful way to allow for more finely divided rate categories, yet not face increasing numbers of parameters to optimise, is to treat λ_j (the rate at the *j*-th site) as randomly drawn from a specified statistical distribution (which will usually have only a few defining parameters). We will also specify that this distribution has mean fixed to one, as we are only interested in relative rates (the absolute rates can be thought of as dictated by the edge lengths in the tree), and all values must be positive (since only positive rates make sense under our evolutionary models). And this is exactly what we were doing in chapter 2. One such distribution is the Γ distribution (see table 2.5 for other examples and their probability density functions). Thus the probabilities of site patterns under this model is still just a sum, $\mathbf{s}_{all} = \Sigma \mathbf{s}_{\lambda j}$, where the λj are randomly drawn from a specified distribution (and in the limit as $c \rightarrow \infty$, this summation becomes an integral if the distribution of rates is continuous). This result is proved in Steel *et al.* (1993) (see also appendix 2.1). Further, under the generalised Kimura 3ST model, this integration can be simplified (by being made at the level of the pathset transformations), or avoided altogether if the distribution of interest has a closed form moment generating function (as the Γ distribution does, see chapter 2). These are the likelihood calculations underlying Waddell and Penny (1995). As mentioned earlier, these studies were made in 1993 (Ziheng Yang can confirm this as we were in communication at that time), and were sent for publication in 1993 by the editors. Due to delays at the publishing house they have yet to appear (although Waddell and Penny 1995 is available as a preprint from the authors). It is useful to call this the 'integrated' approach.

For more general mechanisms of evolution, Yang (1993) resorted to numerical integrations of the form $\mathbf{s}_{all} = \Sigma \mathbf{s}_{ij}$. If this distribution is continuous, then the cost of making this integration by numerical methods is limited only by the degree of precision that is required (do we chop the continuous distribution up into 5 rate classes, 100, 100,000 ...). In Yang (1993), this integration was done to a high degree of accuracy, with the result that he was only able to analyse four taxa.

To avoid the great burden this imposes, the simplest solution is to consider doing a crude numerical integration by taking the continuous distribution and integrating in just a few slices. I suggested this to Ziheng Yang in May 1993 (via e-mail), following some early evaluations I had made under the Cavender model. He replied acknowledging this as a "good idea" that he thought should work, but has not acknowledged this in his publications. In Yang (1994) it is shown that such integrations could be surprisingly accurate with as few as four rate classes, with the integration simply taking an equal proportion of sites in each rate class (for more accuracy it may be desirable to use something like Simpson's rule to determine the proportion of sites in each rate class). This method has been called the discrete Γ approximation, and of course it can take its rate classes from any distribution (e.g. the lognormal, the F etc.) at practically no extra cost.

Because the histogram-like classes of discrete approximations are calculated from a continuous distribution, then no matter how many classes are specified, the only free parameter in our model is the shape parameter(s). The promising performance along side the exact Γ distribution evaluated under Hadamard models, suggests that for say 10 rate classes (therefore order 10 times the computational effort of the i.i.d. model) the approximation to the true distribution becomes very good in many instances. The fact that Yang (1994) obtained an optimal fit of data to model with as few as 4 rate classes is due to the fact that the real data has sampling error and probably does not conform exactly to a Γ model (it's an open question exactly what it conforms to, although some sort of approximately L shaped distribution seems likely).

Even more computationally simple is a finding in the studies for Waddell and Penny (1995) which found that even the simple invariant sites / i.r. model would give a likelihood similar to a model assuming a Γ distribution of rates across sites. Importantly for tree building it is very hard to construct sequences under one of these two models and have the other model find a different optimal tree (exact sequences were predicted with Hadamard conjugations, with some such results shown later in this thesis).

The main distinction that arises with comparing a "discretised" continuous distribution with a continuous distribution, is in the tail of a distribution. Fortunately, as noted in chapter 2, this is an area which we generally don't expect to finely resolve, although it is important in seriously evaluating which (if any) of the standard continuous distributions best fits the data. It is possible to predict some features of the "discretised" models relative to the continuous parent distribution. It is possible to predict the differences that will arise. Given $\mathbf{s}$, a discretised model will suggest the same $\mathbf{r}$ as an exact distribution, but smaller ρ values especially on the longest paths. Going the other way, for a given T and k, the discretised model will predict the same ρ but a smaller $\mathbf{r}$, and so a more spread out $\mathbf{s}$. Assuming that the shape of the distribution is fixed, discretised

models tend to predict shorter pathlengths (and pathset lengths), and hence also shorter edges on trees (when sites are grouped about their mean or median rate). Alternatively if the shape of the distribution is being optimised, there is a trend for the discretised distribution to infer that a distribution was more like i.r. than the corresponding exact calculations (e.g. k will tend to take slightly lower values). These effects may become more prominent as taxa are added. This does not always show up with real data, since the data does not fit the model, and the true distribution is not Γ.

On the other hand, a sensible feature of discretised distributions is that they are in some ways like a truncated distribution, which makes sense in that we do not expect sites to generally be going faster than the neutral rate. Overall it seems likely that ML methods will not require laborious effort to gain substantial robustness to rates across sites, and that even discretised distributions or invariant sites models will bring a great deal of robustness. The best robustness, and perhaps the most interesting information, will probably come with the implementation of independent discrete rate class models like those of Felsenstein (1993) which allow optimisation of site rate classes directly. Such models should pick up features such as distinctly multimodal distributions which need not be apparent with the continuous distributions. As we noted earlier, the addition of the optimisation of another 10 parameters should not slow down the rate of convergence to optima dramatically (at least with 4 - 10 taxa), although the cost in computation will rise linearly with each rate class added.

Olsen (1994) has taken another approach to ML to compensate for a distribution of rates across sites. This approach is more like weighted parsimony (e.g. Farris 1969). It does not attempt to estimate the overall likelihood of the data given the model with rates across sites, but rather evaluates likelihoods under an i.r. and i.i.d. model, then weights the contribution of each site to the overall likelihood. The weights for sites (w_j) may be derived separately, or perhaps in some iterative fashion (with likelihood the iteration will often be on the first tree, due to computational cost). Thus the likelihood is the product over all sites (j) of $s_j w_j$ i.e. $\ln L = \Sigma \ln(w_j s_j)$ (summed over all j). A sensible way to choose these weights is inverse to the intrinsic rate at a site. Given a weighted tree, and a mechanism of change, the estimated rate for site j is given by finding the $\lambda_j s_j$ which returns the highest likelihood (where λ_j is optimised over the range $0 \rightarrow \infty$). Alternatively, a tree independent method such as that of Penny and Hendy (1985) may be useful to rank site rates (after modification for this purpose), (Felsenstein 1981b also offers some useful insights into a logical weighting scheme). This method is certainly quick, and if the weights are reliable, then it probably offers a high degree of robustness to unequal rates across sites. The weights will tend to become more reliable the more sequences there are, a good example being those available in the large RDP rRNA sequence data base (Olsen *et al.* 1992). Conversely, site rate estimates could be misleading if only a few sequences are available, and this could lead to inconsistency of the type in Felsenstein (1978a), or Hendy and Penny (1989), with convergence to the wrong answer even speeded up over i.r. likelihood. This type of model is very much a tree estimation approach, it returns minimal information on the process of molecular evolution.

Yet another approach is to classify sites into rate classes. The disadvantage of this approach is that it can be very hard to do this accurately with few sequences. This approach has been used in likelihood calculations by Barry and Hartigan (1987a) and by Ritland and Clegg (1987) (in both cases the classification was into codon position). Yang (1994) briefly studied using the likelihood of a site pattern on a star tree to classify sites into rate classes. Not surprisingly, this application of the method did not seem as reliable as the 'integrated' approach. We would expect that using likelihood (or just parsimony) to classify sites into rate classes will begin to work well as the number of sequences grows, and we have some moderately reliable way of inferring the underlying tree. Indeed as we noted in chapter 3, the unconstrained i.i.d. model with a different **P** matrix on each edge requires sites to be classified into rate classes so that each set of sites are approximately i.r. The parsimony length of a site on a tree should also provide a good guide to the relative rate of sites (one which may be little inferior to exact likelihood calculations when imposing i.i.d. models on sequences evolving under complex covarion models).

A very useful side effect of classifying sites into rate classes is that it avoids to some extent the 'base composition lag effect' which we noted in chapter 3 which will occur when we model rates across sites with nonhomogeneous substitution processes. That is, the slowest evolving sites have a base composition which changes more slowly than that of the faster evolving sites, even if we assume all sites evolve by a similar rate matrix in each interval on the tree. This non-homogeneity of base composition by rate class effect can be a real problem. We suspect it is resulting in LogDet transforms of all variable sites, and FD tests (Steel *et al.* 1993d) on all variable sites, usually undercompensating for base composition differences (including the earliest eukaryote divergences, and inferring the origins of plastids). The other great advantage of this approach is related to the points developed in section 4.5.4. The reliability of the model can be evaluated for each rate class, and we should bear in mind that for stochastic reasons, or due to suspected systematic errors, we may wish to exclude or down weight the contribution of the higher rate classes to selecting which tree is best. We expect that such an approach will be very important with deeply diverged functional molecules.

Integrated likelihood approaches are applicable to nonhomogeneous models, with rates across sites, and shifting base composition. Yang and Roberts (1995) give an application of such a model. This model specifies a base composition at the root which applies equally to sites in all rate classes, and has a rate matrix of the general form of that of Hasegawa *et al.* (1985) on each of 6 edges in the tree, with each rate class having a scalar of the underlying rate matrix. A number of rate classes are generated by a discretised Γ distribution, which is in turn optimised via the shape parameter k. Evolution of all sites proceeds down each edge of the tree, usually with a transition matrix out of equilibrium with the base composition at an earlier node in all rate classes. The overall likelihood of a site pattern in the data (which is not split up) should then be the sum of the likelihoods in each rate class, i.e. $\ln L = \Sigma s_i \lambda_i$. The number of optimised parameters would be k, the starting base frequencies (three independent values), and how many parameters in the rate matrix are modified on each edge of the tree (this could be up to 5 with the Hasegawa *et al.* 1985 mechanism).

For this type of model, a number of extensions are obvious. The master rate matrix on each edge of the tree can have up to 12 free parameters, and the base composition of the root could be separated up to the point of allowing each rate class to have a different starting base composition (this is logical since if there was prior nonhomogeneous evolution this could be the case). It is even permissible (and perhaps informative) to separate edges into two or more segments, each of which has its own rate matrix. However the model cannot step to the very general i.i.d. model of Barry and Hartigan (1987a) without separating sites into rate classes, because with a general **P** matrix, there is no way to specify the **P** matrix in each rate class as a simple function of a master transition rate matrix from that segment. The trade off is that the general **P** matrix returns less specific information about the substitution process on an edge (although it is not clear how much information will be reliable given sequences of a few thousand bases, and a model which does not match the true, covarion process of evolution).

A possible disadvantage of classifying rates into site classes is that the number of parameters can increase rapidly if one allows complete freedom of the model within each rate class. The alternative is simple; fix edge lengths on one "master weighted tree", and the tree for each rate class has the edge weights of this tree times some constant. It will be interesting to see how well this works with different types of real data. We expect that with diverged sequences, there may appear to be significant support for allowing each rate class to have its own unique parameters. However this could well be an accommodation effect, where by a model which does not explain the data adequately, can find other parameters to fiddle in order to substantially increase goodness-of-fit. With the general i.r. and i.i.d. model (or any model which accommodates shifting base compositions, with the possible exception of the Kimura 3ST model), then with shifting base compositions it is practically impossible to accurately estimate edge lengths as the number of substitutions per site. In these cases, the length of edge *i* may have to be measured as $\ln[\det(\mathbf{P}_i)]$ (natural logarithm of the matrix determinant). This being the case it will be important to evaluate how this method performs, since we already know from chapter 3 that this measure will tend to overestimate edge lengths most when base composition is unequal. This could lead to a distortion, such that the tree for the faster evolving sites, needs some edges to be "extra long" relative to those same edges for the slower evolving sites, even if the relative rates remained fixed (and the process was truly i.i.d.). This in turn suggests that it may be difficult to test whether fast and slow sites have the same substitution probabilities, when base composition is non-stationary, and edges on the tree are relatively long.

5.3.5 ML models where sites change their rate class.

In this section we consider the mechanics and application of a new class of ML models, ones where sites change their intrinsic rates. Such models bring an added touch of realism, and impinge upon the issue of evaluating likelihood under covarion type models. In the first model considered, there is a transition from a functional to a non-functional gene (i.e. a pseudogene) or perhaps the reverse (i.e. turning on a pseudogene). This sort of model is interesting to consider, since in practice a tree of pseudogene sequences is often rooted with a functional gene sequence (e.g. the pseudogene psi-eta is rooted with a functional eta globin gene, Bailey *et al.* 1992). To illustrate we use a simplified model, where the sites in the pseudogene evolve at identical rates, while those in the functional gene fall into three i.r. categories, coincident with 1st, 2nd and 3rd position sites. A useful feature of this model, is that the overall sequence pattern probabilities, $\mathbf{s}_{all}$, can be considered as a finite sum of components. Specifically,

$$\mathbf{s}_{all} = 1/3(\mathbf{s}(1) + \mathbf{s}(2) + \mathbf{s}(3)) \qquad (5.3.3\text{-}1)$$

where $\mathbf{s}(1)$ is just the sequence patterns corresponding to 1st position sites. Since sites in each rate class have a fixed relative rate in all parts of the tree, then $\mathbf{s}(1)$ is the set of sequence pattern probabilities predicted by a single weighted tree (see figure 5.4). The same condition holds for the other two sets of sites. The overall model is shown in figure 5.4.

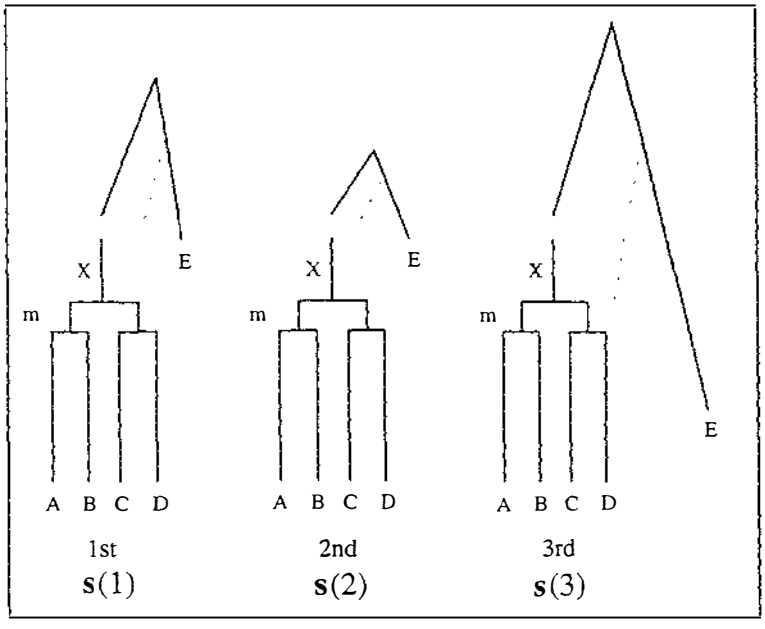


FIGURE 5.4 The three weighted trees that make up the pseudogene-functional gene model. Sequences A-D are pseudogenes, while E is an homologous functional gene outgroup. Special parameters of this model, are X, the time before the common ancestor of the pseudogenes that the functional gene became a pseudogene, Y the length of the edge from the tip of X to E, and λ_1, λ_2, λ_3 the relative rates of coding positions in the functional gene. With only one functional sequence, E, the number of parameters required to fully describe the edge leading to sequence E is just 3, so we cannot solve for X, Y, , λ_2, λ_3 separately. However if we add in another functional sequence (shown by the dotted line), then we have 3 × 3 functional gene edge lengths, and the set of 7 parameters (X, λ_1, λ_2, λ_3 plus the three edge lengths replacing Y) describes them and the sequence patterns they generate. Notice that because we can infer X, it is possible to obtain a refined

estimate of how long it has been since the functional gene turned-off. If we make m small and / or the edge leading to the first divergence amongst the functional genes large, then these trees will be like those in Hendy and Penny (1989) (that is parsimony can become inconsistent in estimating them).

So given that $\mathbf{s}_{all} = 1/3(\mathbf{s}(1) + \mathbf{s}(2) + \mathbf{s}(3))$, it is possible to optimise the likelihood function using equation 5.2.1-2 or 5.2.1-5, as soon as the relationship between T(1), T(2) and T(3) is clarified. In the example in figure 5.4, the 5-taxon trees are identical except for the edge which includes the root. This edge (e_E) in turn can be divided up into two components: the first part after the switch from functional to non-functional (which has a common value on all three site trees, labeled X) and the second part prior to the switch which will be unique to each site tree (but estimated as $Y\lambda_i$, where λ_i is the rate class in the functional gene). If we do not fix the relative rates of sites in the functional gene, this model could have up to 2(t-z)-2 (the edges amongst the pseudogenes), plus 3z (edge lengths amongst the z functional genes) free parameters. Even under the two parameter Poisson model, when *t* is at least 5 there are more distinct patterns in that data than this. If we fix the relative rates of site classes in the functional genes, then the parameters in the model become the 2(t-z)-2 edge lengths amongst the pseudogenes, plus X, plus the 2z-1 edge lengths for the functional first position sites, plus the relative rates of the second and third position sites (λ_2 and λ_3). Since we also know which sites belong in each rate class, it makes sense to keep sites in these classes, and estimate the likelihoods with $\mathbf{s}(i)$ for each one of them (keeping of course the fixed model relations between the weighted trees for each coding position).

An important feature of the model shown in figure 5.4 is that the tree fro each site is potentially in the Hendy-Penny zone of inconsistency (Hendy and Penny 1989), which can mislead parsimony applied to the observed sequences. However, while applying the Hadamard conjugation to each coding position separately will restore consistency of parsimony under the model, applying the Hadamard to $\mathbf{s}_{all}$ need not. It is not appreciated that a parallel situation can exist for ML tree selection (later in this chapter). Here the ML tree selection criterion applied under the specified mechanism of substitution to each coding position will be consistent, but it is potentially inconsistent when the data are grouped as $\mathbf{s}_{all}$, unless the new model described here is used. It is also important to note that while application of ML to each site will recover the tree for each site, it will not give a direct ML estimate of all the parameters in the model (including X).

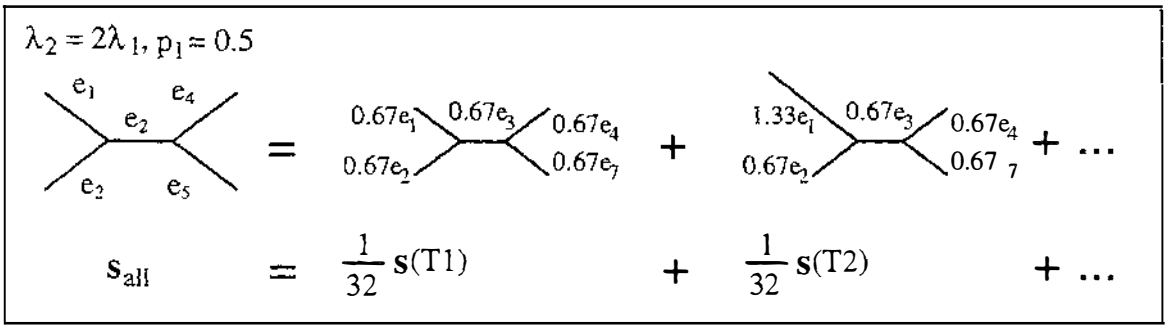


FIGURE 5.5 A depiction of a unequal rates model where sites switch rate class. The underlying model (on the right) has sites falling into two rate categories, $\lambda_2 = 2\lambda_1$, so fixing the mean to 1 gives $\lambda_1 = 1/1.5$ and $\lambda_2 = 2/1.5$. Sites randomly change their rate classes. The sequence pattern probabilities for this model (s_{all}) may be considered a sum of the sequence patterns for 32 different weighted trees (where edges take on all possible permutations of λ_1 and λ_2).

The second model described here is a relative of a covarion model, in it sites change their rate class randomly at each internal node on the tree. For simplicity we consider a model with just two rate classes, $\lambda_1 \neq \lambda_2$. Under this model, s_{all} will again be a sum of different weighted trees. There will be 32 distinct weighted trees, which represent all permutations of the model tree, having either λ_1 or λ_2 on each edge (see figure 5.5). Assuming the true tree to be T_{12} with edge set $[e_1, e_2, e_3, e_4, e_7]$, the first permutation would be a weighted tree with edge weights $\{\lambda_1 e_1, \lambda_1 e_2, \lambda_1 e_3, \lambda_1 e_4, \lambda_1 e_7\}$, the second would be $\{\lambda_2 e_1, \lambda_1 e_2, \lambda_1 e_3, \lambda_1 e_4, \lambda_1 e_7\}$, and so on through all the binomial combinations until we arrive a $\{\lambda_2 e_1, \lambda_2 e_2, \lambda_2 e_3, \lambda_2 e_4, \lambda_2 e_7\}$. If it is specified that the proportion of sites with rate λ_1 is always p_1 (a constant across the tree, implying that when one site changes its rate class another site from the other rate class replaces it), then for this model, s_{all} is a weighted sum of weighted trees, specifically,

$$s_{all} = (p_1)^5\{\lambda_1 e_1, \lambda_1 e_2, \lambda_1 e_3, \lambda_1 e_4, \lambda_1 e_7\} + (p_1)^4(1-p_1)\{\lambda_2 e_1, \lambda_1 e_2, \lambda_1 e_3, \lambda_1 e_4, \lambda_1 e_7\} +$$

$$\ldots + (1-p_1)^5\{\lambda_2 e_1, \lambda_2 e_2, \lambda_2 e_3, \lambda_2 e_4, \lambda_2 e_7\} \qquad (5.3.5.\text{-}1)$$

Further, if we add the constraint that $p_1\lambda_1 + (1-p_1)\lambda_2 = 1$, then the total number of substitutions on each edge of the tree will be equal to $\{e_1, e_2, e_3, e_4, e_7\}$.

It should be useful to study such models which in many ways approximate covarion models. To restrict the rate at which sites change rate classes so that they do not necessarily all change at each node in the tree, it is necessary to add in an extra weighting to each tree, which is proportional to the probability of a site showing the observed number of rate shifts on each tree. If the conjecture in appendix 2.1 is correct, then under this sort of model standard methods will tend to overestimate the distance between taxa, and this overestimation becomes worse as the distance increases. Thus after applying transformations which infer unseen substitutions, there is the possibility of an anti-Felsenstein zone problem of long edges being repelled (see section 5.6) should we attempt to rebuild the tree with standard methods.

When first considering the uses of equation 3.2.2-3 (the prediction of the overall transition matrix on an edge with unequal rates across sites), it seemed that it might be possible to predict

likelihoods of more than two sequences given a stationary distribution of rates across sites, using such transition probabilities in place of those used by Felsenstein (1981a). This is not correct, since there summation needs to be made of the **s** vectors generated by all weighted trees consistent with that distribution (as described in section 5.3.4). However we now conjecture that if the type of **P** matrix described in 3.2.2-3 is used in place of the standard i.r. transition matrix of each edge (and fixing the form and parameters of M_λ over all edges), then what we have is a model of sequence evolution, where sites on edge *m* have rates distributed according to the underlying distribution of M, but every time a node in the tree is crossed, these sites loose their previous rate category, and are reassigned a new one drawn randomly from the underlying distribution of rates across sites. We could increase the rate at which sites changed their rate class by introducing internal nodes along edges. Accordingly with just one node along an edge *m*, $\mathbf{P}_m = M(\mathbf{R}t)$, with two nodes along that edge $\mathbf{P}_m = (M(\mathbf{R}t/2)) \times (M(\mathbf{R}t/2)) = (M(\mathbf{R}t/2))^2$, with *n* nodes $\mathbf{P}_m = (M(\mathbf{R}t/n)^n \rightarrow (\mathbf{I} + \mathbf{R}t/n)^n \rightarrow \exp(\mathbf{R}t)$, as $n \rightarrow \infty$ (where the subscript denotes the number of evenly spaced internal nodes on each edge). Thus if we have sites changing amongst the different rate classes quickly, the model becomes indistinguishable from the standard i.r. model of Felsenstein (1981a), or Barry and Hartigan (1987a). Our next aim in the study of these models is to prove that our interpretation of a model where the transition matrix on each edge is generated according to equation 3.2.2-3 is correct, and then seek to extend the models to allow weighted sums so that only a proportion of sites change their rate class at each internal node. These models are a step towards covarion models, with the main factor missing seeming to be functional correlations between particular sites.

Models which are considering some other factors related to covarion models, but not yet allowing for unequal rates across sites are those of Tillier (1994) and Schöniger and von Haeseler (1994) which model rRNA evolution where paired sites (e.g. pairs of sites forming the "stems") coevolve. All these models are showing intriguing results. It will be important to find out if stable covarion models (i.e. those that don't change the substitution pattern dramatically through time) predict regions where i.i.d. methods of tree estimation are as susceptible to the long edges attract type of problem as parsimony, for example, applied to observed sequences.

5.3.6 Results with ML models that allow distributions of rates across sites.

The following three sections give examples of analyses with one of the newest, and hopefully most robust and possibly statistically efficient tree selection methods. As mentioned earlier Yang (1993) developed conceptually equivalent methods for calculating likelihoods given a continuous distribution of rates across sites. Appropriate citations for this independent development of likelihood methods with continuous distributions of rates across sites are Steel *et al.* (1993c) for a proof of the consistency of these methods, and Waddell and Penny (1995) for their application (this last publication accepted for printing in 1993), or this thesis.

The likelihood calculations illustrated here, use the extended Hadamard conjugations developed in chapter 2 to predict either the 2^{t-1} sequence patterns with 2-state, or 4^{t-1} patterns with 4-state data. From these vectors and their observed counter-part, $\hat{\mathbf{f}}$, the G^2 statistic is calculated. Minimisation of this likelihood statistic, and other goodness of fit statistics, involved simultaneous optimisation of all free parameters using a Newton-type algorithm, followed by at least one application of a conjugate gradients method to check that a local optimum had been found. That there was likely to be just one likelihood minimum per tree, was checked by starting the optimisation process from a variety of randomised starting points (this was also useful for checking that flat spots on the likelihood surface were not causing problems). The most general model considered here has 3 free parameters per edge (the rate of transitions and the two types of transversions on that edge) and the parameters for the distribution of rates across sites (generally one free parameter). The most complex distribution used here is an invariant sites plus a Γ distribution. For this data, the optimal solution under the invariant sites-Γ model was to drive any initial proportion of invariant sites to zero, and allow k to take its optimal value of 0.77.

The opposite of developing ever more complex models, is identifying reliable simple models, which is in practice equally important. This is because the variance of estimates increases with the number of parameters added to a model. There are, however, many ways in which to simplify, for example, the generalised Kimura 3ST model (which has 3 parameters per edge plus a distribution of rates across sites). Here we examine a series of logical simplifications of this model. One step of complexity down from the 3 parameters per edge model, we have 2 parameters per edge free to vary independently of each other, but with the third type of change linked to one of the others by a constant ratio. There are three possibilities here; transition rate optimised on each edge, transversion type one and transversions type 2 must be in a constant ratio but their absolute rate per edge free to vary. Likewise transitions (tr) and transversions type 1 (tv1) may be constrained to a constant ratio on an edge, and the third possibility has transitions and type 2 transversions (tv2) linked. We call these mixed intermediate models, and for figures 5.6, 5.7, and 5.8 label them as *i1*, *i2* and *i3* in the order of their description here. The next simpler model is the non-homogeneous generalised Kimura 2ST model, with a different rate of transitions to transversions on every edge (labeled *2p* in fig 5.6). The homogeneous Kimura 3ST model has one parameter per edge with 2 parameters for the whole model which determine ratio of tr to tvl and tr to tv2 (this model marked as *3ps* in fig. 5.6). The homogeneous Kimura 2ST model, one parameter per edge, 1 ratio of transitions to transversions homogeneous across the

whole tree (*2ps* in fig. 5.6). And lastly the Poisson or Jukes-Cantor model with 1 parameter per edge and all substitutions equally likely. We will use a *C* (e.g. *3PfC*) to denote a model with a clock constraint (i.e. all tips can be drawn equi-distant from a nominal root)(used in figures 5.7 and 5.8).

It does not generally make sense to consider a strict clock when different types of substitutions are varying in rate, so here we confine a clock to the fixed transition / transversion ratio models (the homogeneous models). This is because there is no known biological hypothesis to explain why a non-homogeneous mutation process should regulate its parameter changes to produce an ultrametric. In general we found that imposing a clock caused the calculations to find the optimal point to run considerably faster per tree (e.g. with the six hominoid mtDNA sequences nearly twice as fast). This result makes sense in that there are only about half as many parameters to optimise. It is surprising then that the DNAMLK tends to run considerably slower than its non-clock equivalent DNAML (Felsenstein 1993), although this may be due to the need to consider the (2t-3) times as many rooted as unrooted trees.

The different submodels have slightly different sufficient (fully informative) forms of data (e.g. see section 4.6.1). For example, under the Jukes-Cantor model the patterns AGGG and ATTT are equivalent and could be grouped together when testing the fit of data to model (i.e. pool observed data into equivalent patterns and then make ML optimisations). Either way the lnL statistic will be the same, but the G^2 statistic will be different. Pooling data to the sufficient statistic before applying a goodness-of-fit-test may either hide differences between the observed and predicted data, or alternatively pooling data can make the model more sensitive to violations of its expectations (we suspect it will generally tend to increase sensitivity to deviations from the generalised Kimura 3ST model). Here all calculations of G^2 are made with 4^{t-1} distinct patterns expected under the generalised Kimura 3ST model, regardless of any further constraints on parameters. This allows more direct comparisons of these statistics for the different models without needing to worry about differences in degrees of freedom etc. To directly compare the G^2 of the models used here, with models whose sufficient statistic are the 4^t patterns, would require the fit of all models to be evaluated on the 4^t patterns. Note, the Kimura 3ST submodels can specify a non-equifrequency root base composition in order to improve fit when it is measured on the whole 4^t patterns. However, as section 2.4.1 points out, if fit is measured on the 4^{t-1} distinct patterns that arise under the Kimura 3ST model with equi-frequency base composition, this feature of non-stationary root base composition cannot improve fit. Thus if one of these models does not fit to the 4^{t-1} patterns, it is not expected to fit to the 4^t patterns when allowing for unequal root base composition.

There is no problem in converting these likelihoods to those reported by programs such as DNAML if we assume equifrequency of the root base composition. The log likelihood of the tree model (lnL_T) measured on the sufficient order 4^{t-1} data patterns of this model is $-1/2G^2 + lnL_U = -1/2G^2 + \Sigma\hat{f}_i \ln(\hat{f}_i / c)$ for all observed patterns (see equation 5.2.1-5). For the data sets used here, lnL_U was -3,199.07 for the 4-taxon rRNA data set edited from Lake's (1988) alignment, and for the 6-taxon Horai *et al.* (1992) dataset minus just the deletions, it was -7,844.89. As an example,

the best fitting 3PΓ model for the Horai *et al.* data has a G^2 of 303.20, so lnLT(for the 4^{t-1} data form) = -303.20 /2 + -7,844.89 = -7997.42. Converting this $\ln L_T$ to its value under the 4^t patterns is easy if we assume the model has equifrequency root base composition. The likelihood $\ln L_T$(for 4^t data patterns) is $\Sigma\hat{f}_i \ln(s(T)_i /4)$, = cln(4) + $\Sigma\hat{f}_i \ln(s(T)_i)$ = cln(4) + $\ln L_T$(for the 4^{t-1} data form), which for our example is 4,898ln(1/4) + -7,884.89 = -6,790.07 + -7,884.89 = -14,787.5.

However, optimising the assumed base composition at the root will decrease this number, and the likelihood then is not so easy to calculate. Whether this model is considered realistic depends on a number of things, including whether there is good evidence to show that the base composition is shifting in this particular way (from unequal at the root, towards equifrequency in all taxa). This seems unlikely in most situations. It is however debatable whether this modification can be routinely invoked if it makes a significant difference to the overall fit. This is because while it is unrealistic, is it any more unrealistic than adding a Γ distribution i.i.d. assumption to molecules which we know must be evolving by some sort of covarion model (e.g. the ancient rRNA's, see section 2.7)? Alternatively, altering the root base composition may be looked upon as a way of factoring out the contribution of the gross base composition to the likelihood of the model. (giving a gross approximation to the models of Felsenstein 1984, 1993, and Hasegawa and Kishino 1985, if these are generalised to allow tv1 and tv2 to go at different rates, and when the differences between all taxa are small). If amounts of change are not large, then this seems reasonable. Whether the root base composition is assumed to be in equilibrium, or not, will not (in expectation) affect the parameter values by the generalised Kimura 3P model, as long as all fitting is done on the 4^{t-1} patterns, otherwise differences are expected. Here, all fitting is done on the order 4^{t-1} form of the data, so the results are most applicable to the assumption of equal base frequencies at the root.

5.3.7 ML analysis of four ancient rRNA sequences

Here, four 4-state rRNA sequences aligned by Lake (1988) are studied using likelihood methods which take account of unequal rates across sites. This data set is similar to that used by Navidi *et al.* (1991), subsequently restudied by Goldman (1993a), Yang (1993), and Yang *et al.* (1994). The same data is studied in 2-state form in figure 2.7, and a similar, extended data set edited from the alignments of Gouy and Li (1989) is studied in chapter 3.

A useful feature of this data, is that it nearly meets the approximate guidelines laid down for trusting the asymptotic distribution of goodness-of-fit statistics like G^2 or X^2. Specifically, there are 64 observed patterns in this data, of which 36 have a frequency of 5 or more, 26 have a frequency between 2 and 4, and just 1 has an observed frequency of 1 and 1 of 0. The expected frequencies under the better models are very similar, with 33 having an expected value of 5 or more, 21 an expected value of 3 to 5, and 10 with an expected value of 2 to 3 (for some guides to conditions for reasonable asymptotic approximations, see Stuart and Ord 1990, p. or Read and Cressie 1988). There are also at least two factors which will be altering the expected fit of this data away from asymptotic predictions. The first is that cell sizes are uneven, and the test may well be losing power compared to asymptotic expectations (i.e. G^2 and X^2 are biased towards

being too small). A second factor is expected to be acting in the opposite direction, and this is a partial recovery of expected degrees of freedom when ordinary ML estimators are used to infer the predicted values (e.g. see Stuart and Ord 1990, p. 1171). This means that unless you are using direct multinomial ML estimates of predicted values, the true sampling distribution of G^2 or X^2 tends to higher values than predicted by the usual asymptotic approximation. (An example of a direct multinomial estimator is to predict the value in two cells expected to be equal by their average, while fitting $\mathbf{s}(T)$ to $\hat{\mathbf{s}}$ to minimise G^2 is an example of an ordinary ML estimator). This effect usually drops away with an increasing number of cells; its magnitude with 64 cells of unequal expected size is uncertain. Overall however, the difference in G^2 should hopefully be reasonably well behaved. Later in chapter 6, we show that such an approximation for the Hominoid data is likely to be much less reliable.

Table 5.7 The ML tree and parameters of four 16S-like rRNA sequences.

The taxa are eu = human, Su = *Sulfolobus*, Ha = *Halobacterium*, Es = *Escherichia*.

The optimal tree in all cases was the eocyte tree, (eu, Su). Optimal edge weights of each type as labeled. The 3P + Γ + p_{inv} model converged to the 3P + Γ model, with p_{inv} going to zero.

		3P + Γ	3P + p_{inv}	3P + inv. Gaussian	3P + inv. Gaussian + p_{inv}
G^2		37.37	46.35	37.78	37.30
shape parameter		$k = 0.771$		$d = 0.298$	$d = 0.490$
p_{inv}			0.279		0.088
tr	eu	0.538	0.521	0.704	0.685
	Su	0.100	0.114	0.131	0.128
	(eu, Su)	0.015	0.011	0.025	0.021
	Ha	0.149	0.165	0.196	0.190
	Es	0.276	0.276	0.369	0.354
tv 1	eu	0.100	0.103	0.143	0.131
	Su	0.003	0.014	0.000	0.000
	(eu, Su)	0.000	0.000	0.000	0.000
	Ha	0.063	0.071	0.078	0.078
	Es	0.073	0.092	0.096	0.093
tv 2	eu	0.277	0.293	0.357	0.350
	Su	0.000	0.009	0.000	0.000
	(eu, Su)	0.044	0.045	0.053	0.054
	Ha	0.055	0.063	0.070	0.069
	Es	0.197	0.201	0.263	0.252
Σedge lengths		1.890	1.978	2.486	2.405
tr : tv1 : tv2		1.078: 0.239: 0.573	1.087: 0.280: 0.611	1.426: 0.317: 0.743	1.377: 0.303: 0.725

Table 5.7 shows the results of ML optimisation under the generalised 3P models. The optimal tree in each case was the eocyte tree. The Γ model returned a k value of 0.771, which is quite extreme considering these sequences have been edited to a high degree to remove any regions of ambiguous alignment, which also tend to be the more rapidly evolving regions of rRNA. The edge lengths are similar to those shown in figure 2.10, which are estimated with the Hadamard conjugation going from $\mathbf{s}$ to γ. There is again clear evidence of two long external edges (to eukaryotes and eubacteria), two short external edges (to halobacteria and 'eocytes') and a very short internal edge. The internal edge has quite uneven levels of support from transitions

(very low), transversions type 1 (nonexistent) and the most substantial support from transversions type 2 (= 0.044).

The invariant sites / i.r. model did not fit quite as well as the Γ model, but gave very similar edge length estimates. These tended to be higher than under the Γ model, on all edges except the longest edge and on the internal edge based on transitional changes (the overall increase was 4.7%). The mixed invariant sites / Γ model showed an interesting behaviour. No matter where in the parameter space it was started from, it converged to the Γ model, always driving the estimated proportion of invariant sites to zero. The program was carefully checked in a variety of ways including by submitting it with synthetic data (with some random error introduced) which was generated by another program under an invariant sites / Γ model. The behaviour appears genuine, and is consistent with this Γ model indicating that there are already quite enough sites inferred to be unvaried. Similar behaviour was observed with this model applied to hominoid sequences in section 5.3.8; there appears to be some antagonism between these two parameters with some data sets.

The inverse Gaussian model fitted nearly as well as the Γ distributed model. The inferred coefficient of variation (standard deviation of the site rates divided by their mean rate) for these two models is, however, quite different. For the Γ model it is $k^{-0.5}$ = 1.14, while for the inverse Gaussian model it is a substantial $d^{-0.5}$ = 1.83 (i.e. 1 over the square root of the shape parameter). This large difference is due to a flat tails effect (see section 2.3.3), and is also reflected in the substantially larger edge lengths, which have a total bigger than under the Γ model by 31.5%. A look at the relative rates of transitions to transversions, shows that the inverse Gaussian model was predicting slightly more multiple hits amongst the transitions (the 3P Γ tr / tv ratio is 2.65, while under the inverse Gaussian this increases to 2.69). From these results it seems likely that a lognormal, Weibill or F distribution would also fit this data well, and because of their even flatter tails would probably be predicting even larger edge lengths (perhaps up to 100% bigger again, given their coefficients of variation when the overall shape is similar, see section 2.3.3). They would probably also increase the estimated tr to tv ratio, although at a guess, probably to no more than 3.0 with this data and model. This gives some idea of how difficult it is to estimate meaningful confidence intervals in this sort of situation of highly diverged data, with long unbranched edges.

In contrast to the Γ model, the mixture of an inverse Gaussian distribution, plus a proportion of invariant sites improved the fit by about 0.48 G^2 units. While this is not significant it does make sense in that the inverse Gaussian, unlike the Γ, does not have a high density function all the way to zero, so that p_{inv} is more likely to have a role to play. The properties of this model were intermediate between those of the p_{inv} / i.r. model and the inverse Gaussian model, with the total sum of edge lengths being reduced to 27.2% larger than under the Γ model. Interestingly, there is a wide range of parameter values over which the invariant sites-inverse Gaussian model fits within about 4 G^2 units of its optimum. For example, with d = 4 and p_{inv} = 0.240 the overall

G^2 is 41.73, while the sum of edge lengths has dropped to just 8.2% larger than under the Γ model, with the estimated tr/tv ratio reduced to 2.51.

Overall the behaviour of ML models with different types of distributions of rates across sites are consistent with predictions in chapters 2 and 3 (especially sections 2.3.3 and 3.5.1). It must be accepted that beyond the usual confidence intervals we must allow for potentially substantial fluctuations in estimates due to other equally well fitting models making different estimates. The absolute sizes of edge lengths are particularly hard to pin down, and this effect is most marked on the long edges as table 5.7 shows. If the length of these edges is to be more accurately estimated, it seems likely that data editing will play the primary role (e.g. see section 4.5.3). The main volatility in the tr / tv ratio appeared to be an effect of flat tails. Since it will be very hard to gauge whether there is a small percentage of rapidly evolving sites with only a few taxa, we should be cautious in trusting our estimates of this ratio from ancient divergences. It would be interesting to plot the estimated tr / tv ratio for progressively more related species, with the expectation being that it will apparently decrease. On the other hand it would be unwise to assume that all transitions are going at about 2.5 times the rate of transversions. Table 3.8 strongly suggests that the most conservative substitutions in this data are transitions. Perhaps we need to consider that the near neutral substitutions are dominated by transitions (and these are what are inferred in comparisons of more closely related taxa), while the more conservative changes (which must be due to a covarion mechanism, see section 2.7) are dominated by quite different mechanisms of evolution which require further study (some possibilities are suggested in section 3.8.1).

The total number of changes per site in this data is certainly large when unequal rates across sites are taken into account (the models returning values in the range of 1.9 to 2.5 substitutions per site). Interestingly, while distances would be experiencing quite large sampling errors at this point (verified by simulations with 'Montreal' a tree selection simulation program by Dr. Paul Lewis), likelihood is still predicting a clear and significant resolution of the short internal edge in favour of the eocyte tree. (the increase in G^2 on the best fitting models assuming the archaebacterial or halobacterial trees is $\approx$ 10 units, while it is slightly more to the star tree). Simulations by Gaut and Lewis (1995) suggest the internal branch length test is reliable when the model fits reasonably well, with α = 5% level when the increase in G^2 is 3.84. This strong support for just one tree is in clear contrast to the analysis of just transversions when unequal rates across sites are considered (see figure 2.7, and also figure 5.23, and see also chapter 6). This is a little surprising given the results in chapter 3, and the claims of Olsen and Woese (1989) which tend to suggest it is the transversions which are principally supporting the eocyte tree. Table 5.7 shows that, at least for this small data set, type 1 transversions do not support the eocyte tree at all. At one level, these results can be taken to indicate the higher resolving power that 4-state data has. At another, as yet not well understood level, they indicate the multitude of systematic biases which are expected to influence analyses of ancient divergences. Clearly the statistical issues of sampling error, pale in comparison to a thorough understanding of systematic biases. It will be important to further dissect these issues, especially as applied to this data. This

is a vast field of study which this thesis can only touch upon. Some would even say that understanding the causes of such systematic errors needs to be resolved in order for phylogenetics to be considered a mature science.

In the next sections we look in more detail at the behaviour of key parameters in evolutionary models. For this we focus upon the two most commonly used approximations to unequal rates across sites: the invariant sites / i.r. model, and the Γ distribution of site rates.

5.3.8 Properties of parameter estimates under URAS ML models

We now look at how the different types of submodel compared on this data. The figure 5.6 clearly indicates the huge difference in fit (measured on the x-axis as the G^2 statistic) between models which incorporate a distribution of rates across sites (figure 5.6a and b), versus those which do not (figure 5.6c). To a first order of approximation, the Γ and the invariant sites distribution fit much better than the i.r. distributed models (on average by 90 to $100G^2$ units). Viewed from this perspective they are very similar in fit, although by an asymptotic χ^2 distribution, the Γ model would appear to be fitting substantially better. An approximate test is to consider the invariant sites / Γ model to be the super model and dropping one parameter generates the Γ model (0 difference) or the invariant sites model (which is worse by 8.98, that is significant at the $\alpha = 0.01$ level on a χ^2 distribution with d.f. = 1). Another difference between models (not shown in the figures) is that the i.r. Jukes-Cantor model also fitted about 80 G^2 units worse than the models allowing for unequal rates across sites (i.r. / J-C $G^2 = 251.61$, p_{inv} / J-C $G^2 = 169.06$ ($p_{inv} = 0.275$), Γ / J-C $G^2 = 169.37$, ($k = 0.837$)), but about 90 G^2 units worse than any model which allowed transitions and transversions different rates. While allowing unequal rates across sites tends to make the biggest difference in edge lengths, allowing for transversions to go slower than transitions often makes a larger difference in overall fit (a point noted also by Yang *et al.* 1994). Also surprising is how under the Jukes-Cantor mechanism, the invariant sites model and the Γ model have reversed their relative goodness-of-fit, showing that unless model assumptions are quite realistic it may be unwise to believe the finer details of differences in models. Note how the estimate of p_{inv} under this simplest of mechanisms was very similar to the most general model (see table 5.7), while the estimate of k has shifted more noticeably. We return to discuss this feature later.

The models which allow the rates of change across sites to follow the Γ distribution, here fit better than the equivalent model using the invariant sites method (compare figures 5.6a and b). The difference appears to be significant by an Akaike type lnLR test of the best fitting model in each class (the difference in G^2 between the very best model of each type is 8.97, which asymptotically should be distributed as chi-square with 1.d.f., Sakamoto, *et al.* 1986, Miller 1990). The models with more parameters generally fit better than those with fewer, but the importance of allowing a separate overall rate for the 3 main classes of changes (tr, tv1, tv2) is indicated by the better fit of the $5 + 2 + 1 = 8$ parameter homogeneous 3ST model (fig. 5.6 a-c, marked as *3ps*) vs the $2 \times 5 + 1 = 11$ parameter 2ST non-homogeneous model (marked as *2p*).

The distinctly higher frequency of type 2 transversions (e.g. see table 5.7) also shows up distinctly in the spectra of figure 2.10. That the mixed intermediate model 2 (*i3*) fits worst of its type, suggests that transitions and type 1 transversions have the most significantly different processes of change in this data. This may, however, also be a reflection that these types of changes are more common in the data (i.e. the changes on the right edge and leading diagonal of fig. 2.10) and the statistic test is therefore more sensitive to their differences.

Some other general trends are also apparent. Figures 5.6a and b suggest that the separation between trees is generally becoming greater as the models approximate the true model more closely (the y = -x trend in the points). This trend is not evident in 5.6c with the i.r. model. We are not sure if the difference in fit between the two non-optimal binary trees is random, or whether it is reflecting some particular bias in each model (this feature is the basis of a test described in chapter 6).

Figure 5.6d indicates that the optimised proportion of invariant sites tends to be larger on the worse fitting trees (the points are p_{inv} under a non-optimal tree minus p_{inv} of the optimal tree). This parameter, like branch lengths, can "be called upon" to help explain the changes which seem to be too frequent under a specific tree model (such as those parsimony patterns supporting the eocyte tree, when the model incorporates the halobacterial tree). In doing so p_{inv}, like edge lengths, tends to be larger under the worse fitting trees. Sometimes with this sort of plot, there appears to be a predominant y = x trend, although often with contradiction (here from the model 2Pf). A similar trend holds with the shape of the Γ distribution (figure 5.6e, where Δk equals k of the optimal tree minus k of a nonoptimal tree) which tends to be more extreme (i.e. implies more unequal rates across sites) on the less well fitting trees. This trend is like that shown in figure 5.3, and it suggests that it is necessary to have confidence in both parts of the model (tree and mechanism of evolution) in order to be sure of accurate estimates of these parameters. While the differences are not so large here, ML modeling of 2-state (R/Y) data for sets of five early eukaryotes from the 16S-like rRNA data of figure 3.12, for example, sees differences as large as $k = 0.65$ on an optimal tree, falling to $k = 0.25$ on a non-optimal tree. As we see more clearly later, both k and p_{inv} also tend to be called upon to help explain excess multiple hits when the other parts of the mechanism don't do this so well (often seen in a comparison of more to less general models). They can do this in a fairly crude way because as the distribution of rates across sites becomes more extreme, there is an increasing probability of the patterns showing parallelisms, convergences, and 3 or more states.

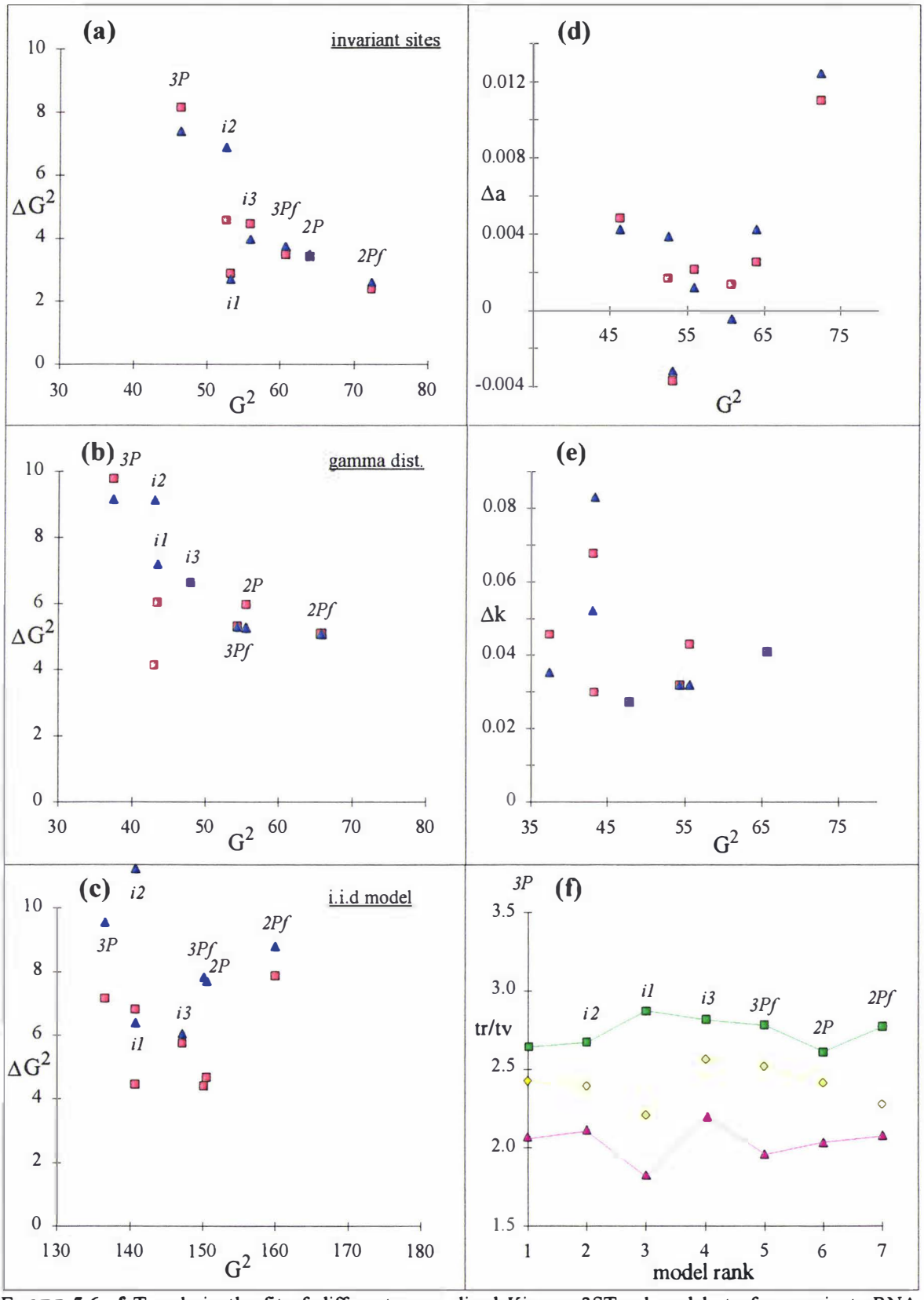


FIGURE **5.6a-f** Trends in the fit of different generalised Kimura 3ST submodels to four ancient rRNA sequences. **5.6a** On the x-axis G^2 for the eocyte tree measured over various submodels (labels given in section 5.3.6), while the y-axis measures how much worse the fit of the archaebacteria tree (squares) and the halobacteria tree (triangles) are. The data are the 4-state rRNA sequences used for the Hadamard conjugation, as described in figure 2.10. Notice that all values are positive (therefore the eocyte tree is

always the best fitting). **5.6b** The same as in 5.6a, except for models with the shape parameter for the Γ distribution being optimised (invariant sites set to zero). **5.6c** The same as 5.6a, but all sites assuming identical rates (i.r.). **5.6d** The same models as in figure 5.6a, with the y-axis measuring the difference in the optimised proportion of invariant sites on the two worse fitting trees. **5.6e** As for (d) but with Γ distributed models measuring how much lower the value of the shape parameter (*k*) was for the other two trees. **5.6f** The inferred transition to transversion ratio for the different models. The Γ models shown in 5.6a are shown as squares (ranked in order from best to worst fitting according to the Γ model), the pinv models are marked with diamonds, and the .i.r. models are marked with triangles. (The tr/tv ratio is calculated as the sum of all transition edge lengths divided by the sum of all transversion edge lengths).

Like the shape of the distribution of rates across sites, the inferred transition to transversion ratio may also vary systematically with the type of model being used. Figure 5.6f shows that in general the tr/tv ratio was higher for models incorporating the Γ distribution (yellow), than for i.r. models (purple). The invariant sites models inferred even higher tr/tv ratios (green). The intermediate models (*i* models in the figure) showed the most volatility in this ratio, possibly because they were each forced to link a different pair of types of change, and some of these match-ups called for altering the tr / tv ratio in order to best fit observed and expected site pattern frequencies. The more general models, here at least, did not give the highest ratio of tr/tv (for the set of models assuming the same distribution of rates across sites). This trend is not without contradiction.

All of the models that allow for unequal rates across sites, and a different rate for transitions to transversions, fit the data adequately when their G^2 statistics are compared with expected asymptotic chi-squared distributions (e.g. the $\alpha = 0.05$ level for the worst fitting invariant sites model with the homogeneous 2P mechanism of change is $\chi^2_{0.05}$ (56 d.f.) = 74.47, while the observed fit is slightly better than this). In chapter 6, identifying which of the adequately fitting models appears to be the most "ideal model" (i.e. fewest parameters to adequately explain the data) is addressed, as is the issue of reliable assessment of the goodness-of-fit when faced with potentially sparse data. It is interesting that the data fits any model well, given evidence that some sites supporting the eocyte tree are due to bias in Lake's particular alignment strategy, and also evidence of an atypical "signal" apparently coming from the halophilic bacteria in relation to the other species in the methanogen group (e.g. Olsen and Woese 1989, and consistent with our analyses of other data sets, see figures 5.1 and especially 3.11 and 3.12).

If this is the case, it suggests that even a good fit between data and model can only be validated after considering the reliability of the alignment. Such a result must also make us very wary in our interpretation of integrated alignment-tree building algorithms such as that of Hein (1990), which could confound both steps of data analysis, and make testing of fit of data to model very difficult. Fortunately the secondary structure in rRNA allows for some true objectively in its alignment. This feature of a fairly close fit of data to model (not necessarily within expected sampling error, but close) is also seen in similar data by Goldman (1993a). Trees built by Yang (1993, 1994) and Yang and Roberts (1995), based on other types of ML model analysing similar data also favour the eocyte tree. Given the analyses in chapter 3 which tended

to give clearest support for the archaebacterial tree (see also Olsen 1987 on this issue), this raises the question of how reliable any i.i.d. model can be for anciently diverged data with sparse species sampling? That is not to say that things will always go wrong, but it is doubtful where standard goodness-of-fit statistics can be relied upon to diagnose this.

5.3.9 ML analysis of Hominoid mtDNA.

Next are presented the analyses that were preparation for the manuscript of Waddell and Penny (1995) which was accepted by the editors for publishing in August of 1993. At the time there were no published methods of making ML calculations assuming a continuous distribution of rates across sites, and a prime purpose of exploring a variety of different models was to better understand the behaviour of these models. Since these calculations were done, Yang (1993), and Yang *et al.* (1994) have published papers on ML models with a continuous distribution of rates across sites, while we published a mathematical proof of the Γ distributed models used in this thesis in Steel *et al.* (1993)(see also appendix 2.2), and were told by the editors to anticipate imminent publication of Waddell and Penny (1995). The reasons for the ongoing delays at Clarendon Press are not clear, however Waddell and Penny is available as a preprint from the authors. In keeping with much of the rest of the thesis results are presented in figure form where the results may be best appreciated. The data were the six aligned 5kb mtDNA sequences of Horai *et al.* (1992), unedited except for the removal of some short deletions in non-coding regions.

The aim of this section is to evaluate some general features of ML models as they might be expected to be applied to moderately divergent sequences. By not editing the mtDNA into rate classes or other "process" partitions (Bull *et al.* 1993) we expect to exaggerate some of these effects. However, this exaggeration is desirable as it gives us some expectation of the sort of "fluctuations" which should be expected in more anciently diverged sequences which have been edited into somewhat homogeneous classes (such as using mtDNA amino acid sequences to infer mammalian order phylogeny and divergence times, e.g. Cao *et al.* 1994) but where the model is clearly an approximation of a much more complex process.

Overall there were 3 main groupings of models as judged by fit to the data (see the x-axis of figure 5.7c). The best models were those which allow for variation of rates across sites and a different transition to transversion ratio, with those modeling a proportion of invariant sites superior to their equivalent except for assuming a Γ distribution of rates (these models have a G^2 statistic in the range of $\approx$ 280-390). The inferred proportion of variable sites (about 40%) is slightly more than the proportion of sites which are third position or non-coding (about 30%), and is probably due to changes in the tRNA genes and some of the less conserved amino acid substitutions. The next set of models, with G^2 statistics on average about 500 G^2 units worse than their counterparts, are identical rate models allowing for distinct transition versus transversion rates. Lastly, there are those models which do not differentiate transitions from transversions; these have very poor G^2 statistics. Amongst these models the gap between the i.r. and the unequal distribution of rates across sites models has closed up considerably. We attempted to optimise

both a proportion of invariant sites, and a Γ distribution for the remaining variable sites. However, on this data as with the examples in the preceding section, these two parameters behaved in an antagonistic manner, with p_{inv} taking its optimal value, and k being driven to a value equivalent to identical rates ($k > 1000$). One interpretation of this result is that the invariant site model was better describing an apparent bimodality in the site rates, between predominantly third positions, versus most other sites. Perhaps the addition of a partial Γ distribution was diluting this bimodality, making for a worse fit to the data, or perhaps the invariant sites model was already predicting an excess of the patterns which the Γ model might otherwise help to explain. (Checks were made, as for the separate program showing similar results with the rRNA sequences (see section 5.3.6). Manually setting k to slightly larger values when p_{inv} was optimal showed a smoothly decreasing fit of data to model, consistent with this being a true minimum).

Next, the relationships and properties of parameter estimates in these models are considered. First to be considered is the correlation between goodness-of-fit of the model and the inferred distribution of rates across sites. More clearly than in figure 5.6, figures 5.7a and b show clear correlations between the goodness-of-fit of a model, and the optimal value describing a distribution of rates across sites. The tendency is for worse fitting models to imply a more unequal distribution of rates across sites. The plots in figures 5.8a and b also emphasise a general similarity in the behaviour of equivalent models, except that a Γ distribution replaces the assumption of a proportion of invariant sites. The relative rank of models remains the same, while differences in G^2 between equivalent models (but for the distribution of rates across sites) is also impressive. This further strengthens our argument made throughout this thesis that the invariant sites models can closely approximate other distributions of rates across sites.

While the trend of worse fitting models tending to predict more unequal rates across sites was clear amongst the more closely related mechanisms of substitution, and also between the fit of trees within a data set (e.g. figure 5.3), it did not apply to four models not shown in figure 5.7a or b. These are the very poorly fitting Jukes-Cantor or Poisson based models (seen in figure 5.6c at the far right). These models returned values of p_{inv} of 0.5391 and 0.5395 for the invariant sites models and k of 0.5662 and 0.5678 (the second number of each pair being for the clock constrained models). As shown later in this section, both these values are significantly different from the values they take when a separate rate of transitions to transversions is allowed. This finding shows how important it is to have major features of the models correct in order to be able to rely on parameter estimates, and associated confidence intervals. Clearly given the approximate nature of all i.i.d. models, our confidence in parameter estimates should be discounted in proportion to how ancient and diverged the data are.

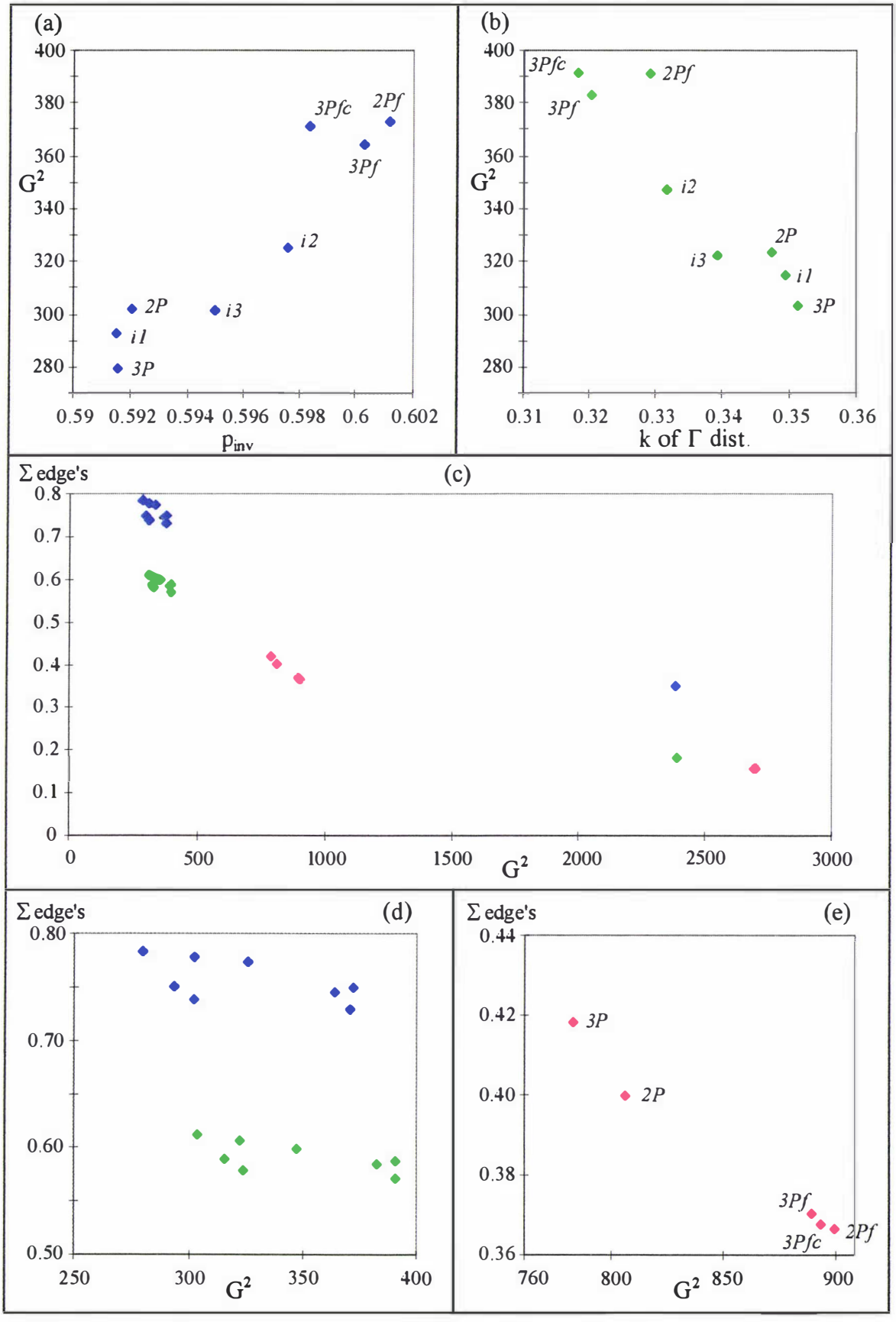


FIGURE **5.7a-e** Relationships between parameter values and goodness-of-fit for a variety of submodels of the extended Kimura 3P set of models (blue are invariant sites models green are gamma distributed models, red are i.r. models, and with labels as per the text of section 3.3.6 shown in (a), (b) and (e) respectively). **5.7a** A plot of the optimal proportion of invariant sites versus fit of model (fit would usually be plotted on (continued next page)

the x-axis but we have plotted it vertically to emphasis the symmetry of this figure with 5.7b). **5.7b**, as for 5.7a except the shape parameter k is plotted against fit. **5.7c** Fit plotted against sum of edge lengths. 5.7d-e Close up views of fit versus edge weights for the invariant sites and Γ models. The trends shown in each case are discussed in the text.

Also of general interest is the relationship between goodness-of-fit of a model and edge lengths in the tree (measured in the expected number of substitutions per site). This is shown in figure 5.6c. The general trend for this data is clear; the better fitting models are predicting larger amounts of change. The invariant sites models, for example, are predicting over twice as much change as the i.r. models, even though this change has occurred over fewer sites (about 60% fewer sites!). Indeed the invariant sites model is suggesting a substantial $0.75/(1-p_{inv}) = 0.75/0.4 = 1.875$ substitutions per variable site! Notice also that the invariant sites models infer substantially more substitution has occurred than the corresponding Γ distribution models. This shows that we need to be confident in the true distribution of rates across sites, in order to be able to estimate accurately how much change has occurred.

It is interesting to consider more closely the relationship of the sum of edge lengths versus G^2 in the models allowing unequal rates across sites, as shown in figure 5.7d. The general trend of longer edges with better fitting models generally holds, but there are exceptions. Again we notice the striking similarity of rank of the invariant sites, versus the Γ distributed models, and it certainly seems they are behaving in a similar manner. In contrast, amongst the i.r. models there is some redistribution of the rank of the edge lengths of model type (the three intermediate models are not calculated in this instance). It will be interesting to see how general the correlation of increasing edge length with respect to goodness-of-fit is when measured across a wide variety of models and data types.

Next we consider the reliability with which transition to transversion ratios are estimated. This ratio is calculated as the sum of all edge lengths measured in transitions, divided by the sum of all edge lengths measured in transversions (results are shown in figure 5.8a). A number of trends are apparent. Generally, the better fitting models which allow for unequal rates across sites infer a higher tr / tv ratio than the i.r. models. However, the trend within the models allowing unequal rates across sites is clearly the opposite. In addition there is a clear distinction, with the Γ models here favouring higher tr / tv ratios than the corresponding invariant sites model. A possible cause of the tendency towards lower tr / tv ratios amongst the best fitting unequal rates across sites models, is that the unequalness amongst transversional changes (i.e. between tv1 and tv2 changes) has lead to the rarer type 2 changes having more leverage upon the model, thus drawing the tr / tv ratio towards explaining these rarer changes better, and so pushing the tr / tv ratio up. By leverage we mean the effect that a few outliers can cause a squares weighted method (for example) to give them more weight in determining model parameters than the bulk of the better fitting observations. The G^2 statistic is closely related to the X^2 statistic (see Read and Cressie 1988 for example), and the latter statistic (as its name suggests) is a weighted sum of squares method. This of course only explains those models where there is not a separate rate for type one and two transversions. The second effect which is apparently pushing up

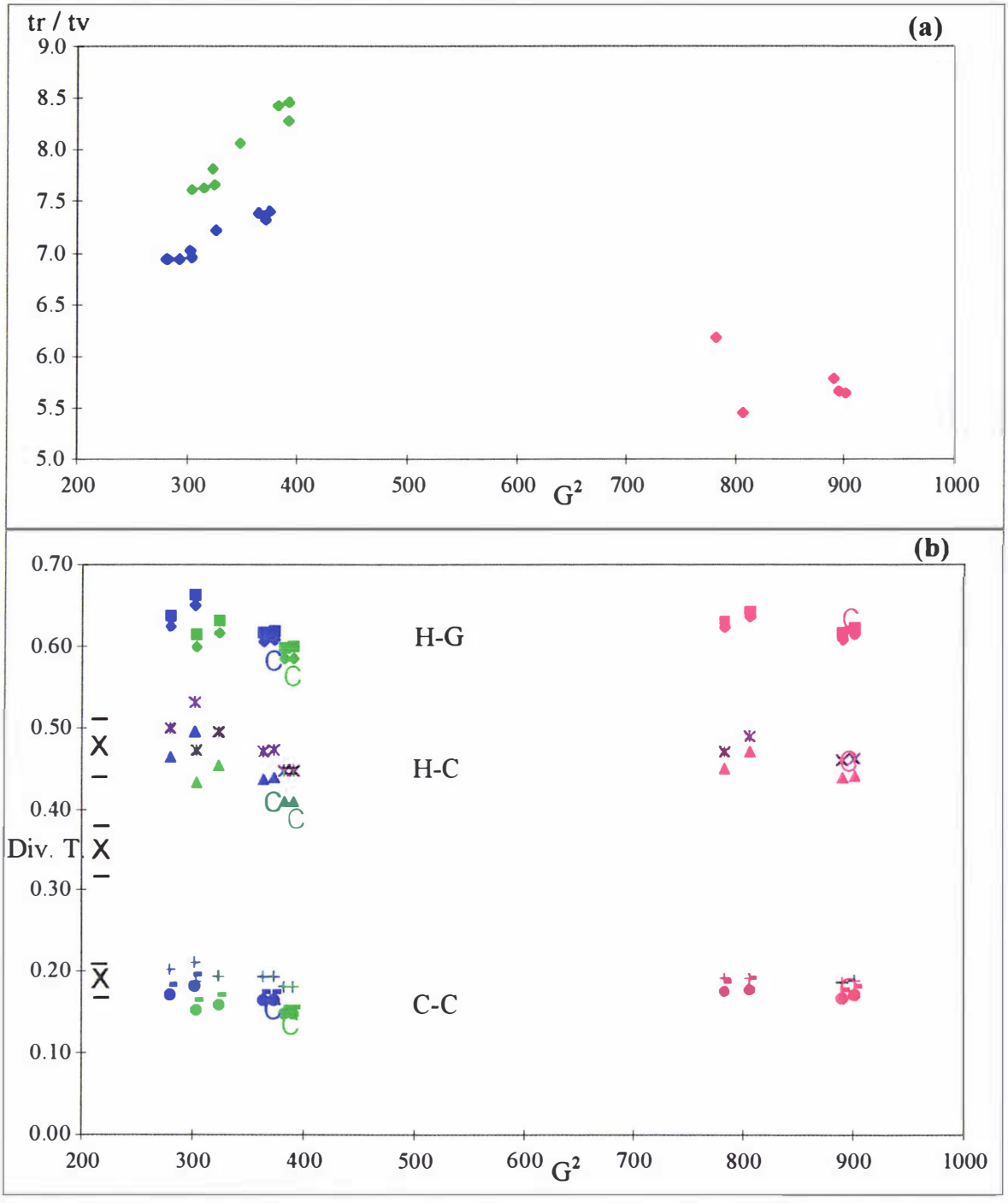


FIGURE **5.8a** A plot of models goodness-of-fit vs inferred transition to transversion ratio (models are the same as those in figure 5.7 excepting the 1P models). **5.8b** Plot of model fit vs estimated sequence divergence times (relative to orangutan divergence). The highest set of clusters (squares and diamonds) are human-gorilla divergence, the next lower are human chimp (stars and triangles), and the lowest are common-pygmy chimp (crosses, balls and bars). Again the models are the same as those in 5.8a, but each divergence time may be estimated in different ways. The chimp sequences showed signs of having evolved more slowly than either human, gorilla or orangutan. Squares and stars represent divergence times estimated ignoring edge to chimps, diamonds and triangles represent their inclusion in averages of edge lengths. The balls are average of two chimp edges to edges leading back to orangutan split, bars are taking more divergent of two chimp lineages only, and crosses are estimating human-chimp split (excluding chimp sequences), then measuring chimp divergence time relative to this split rather than all the way back to (continued next page)

orangutan. The means and standard deviations of all the different human-gorilla divergence time estimates is 0.61 (s.d. = 0.021), for human-chimp it is 0.45 (s.e. 0.029), while for chimp-chimp it is 0.17 (s.d. = 0.016). The large crosses and bars to the left are the estimated divergence times (plus or minus 2 s.e.) using just the observed amino acid differences between the Great apes and humans, as reported in Horai *et al.* 1995).

the tr / tv ratio in the worse fitting models, is an unequal rate of transitions to transversions on different edges in the tree. (At least this is how the generalised 3P model sees it; it is not to suggest there really is a different ratio of transition to transversional changes on each edge, this could itself be an artifact of an inadequacy in the model in relation to the true evolution of these sequences). All in all, then, there is some doubt as to which models are best inferring the transition to transversion ratio. Indeed it is quite possible that the true tr / tv ratio is higher again, and this might only come out by comparing many closely related sequences, or having a model which better accounts for all the relevant factors (including exact distribution of rates across sites, a more exact description of the relative substitution rates, codon usage, functional correlations, etc.). Again this shows that one needs to cautious about the reliability of model estimates as the total amount of divergence becomes non-trivial. It is obvious that we are very much relying upon our models to "correct" for unseen multiple substitutions, and if there is uncertainty in the exact model, this can add a considerable additional uncertainty to such estimates.

Figure 5.8b shows results of much interest to the studies in Waddell and Penny (1995), and a major impetus for developing and applying maximum likelihood models allowing for a continuous distribution of rates across sites. The aim is to estimate relative divergence times of the African apes and humans from coding regions, in this case principally mtDNA. An a priori expectation was that these extended models would infer more recent divergence times, since we anticipated underestimation of the deeper edges in the tree, especially the edge grouping the African hominoids. Such a prediction comes directly from Golding (1983) who showed that not taking into account unequal rates across sites could lead to a profound under estimate of the larger distances (which are paths through molecular-clock constrained trees). In actuality figure 5.8b shows a more complex situation. The addition of a distribution of rates across sites did indeed result in slightly more recent divergence times under all the clock models. However, the general trend was for little change from the i.r. models, with an opposite trend amongst the better fitting models (i.e. towards older divergence dates). Clearly something may be amiss.

That the model-based estimates in figure 5.8b are systematically biased is suggested by the more recent publication of the complete mtDNA sequences of the great apes and humans (Horai *et al.* 1995). Using just the non-synonymous substitutions which showed relatively little evidence of multiple hits, and measuring relative observed path lengths (which do not violate a "clock" amongst the African apes and humans), we arrive at the divergence dates marked by the crosses in figure 5.8b. Estimating the standard errors of these estimates using methods discussed in chapter 6 (see also Waddell and Penny 1995), shows that our model based estimates of the two older divergences are clearly distinct, although the chimp-chimp relative divergence is in good agreement (a trend consistent with a strong underestimation of the lengths of the longer paths in

the tree). At present we cannot explain these trends, they may be due to underestimation of multiple changes amongst the most rapidly evolving sites. This could be alleviated by separate analysis of the main classes of sites, especially separate treatment of first, second and third position coding sites, and the tRNA sites (as identified in Horai *et al.* 1992). It might also reduce bias to analyse at least some genes separately to others, as the study of Cao *et al.* (1994) suggests.

In addition, the average divergence data drops by 10-20% (mostly for the CP divergence time) if the relative divergence time since the orangutan divergence is estimated using a average of all the African taxa's edge lengths i.e. (C+P)/2 +CP+H)/2 +HCP+G)/2 +HCPG) rather than just the backbone lineage (e.g. for the HCP divergence it would ignore the gorilla lineage) as the total time back to the orangutan node (note this makes no difference in the HCPG divergence time, since all African taxa are used there anyhow). Using the orangutan edge as well to estimate the branching point further drops the divergence data more (since the orangutan edge length tends to be large). We did not do this (except under the clock models) as we felt wary of systematic biases in such long edges (but since the clock appears to hold, this may be unfounded). It is important to appreciate that just which lineages are used to estimate divergence time gives rise to a fluctuation in relative divergence dates which can easily be of the order of 10-20%. Different possibilities for estimating divergence dates include: using all taxa (which for ML is best integrated statistically as a clock model): just the taxa in the branch leading from the calibration point down to the taxa whose relative divergence time is being estimated (as done in figure 5.8 due to concern over generation times and possible model systematic biases in the more distant relatives): only the back bone lineage (i.e. an average of all taxa descendant from the node of interest, divided by this number plus just the edges from this node back to the calibration point, again useful if earlier taxa are suspected of showing any biases): or any of the previous methods, minus the edges leading to specific taxa which are suspected of misbehavior (e.g. in this case we sometimes excluded chimps as Horai *et al.* 1992 suspected they may be going slower than average, although clock tests were marginal either way, depending on the test and model assumptions (analyses not shown)).

Most importantly, these results alert us to the difficulties that need to be resolved when relying upon models to take into account multiple substitutions prior to measuring relative edge or path lengths. A good example where this issue will have to be considered, is in inferring the divergence times of mammalian and bird orders in attempting to answer the old controversy of which, if any, of the orders of these groups had evolved prior to the demise of the wingless dinosaurs at the end of the Cretaceous. Because this is a substantially older divergence than those involved here (going back at least three times as far), it will be important to test for various types of systematic error before accepting claims either way.

The relationship between fit and the shape parameter (k) in the 3P Γ model, is a smooth parabola like curve as shown in Figure 5.9a. This smooth shape is very much like that of p_{inv} verses fit (e.g. figure 3.10), or transition to transversion ratio(s) versus fit as seen with this and other ML models (unpublished analyses using Kimura 3ST models and the DNAML program).

This suggests all these parameters should be reasonably well behaved, (being without any apparent discontinuities of multiple optima), especially with regards to constructing confidence intervals about them.

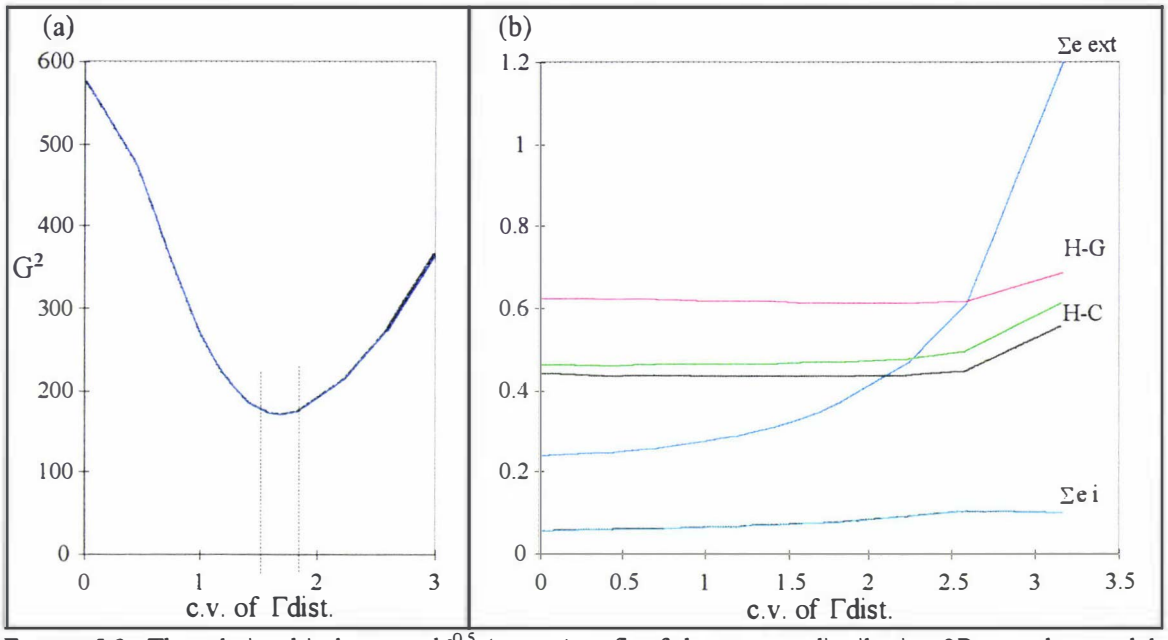


FIGURE 5.9a The relationship between $k^{-0.5}$ (or c.v.) vs fit of the gamma distribution 3P per edge model, for the same 5 species and data used in figure 5.4 (the species being, human, common chimp, gorilla, orangutan, and siamang). Overall the relationship is smooth and approximately parabolic. The vertical dotted lines indicate the points at which the model optimised but for the shape parameter k is worse by 3.84 G^2 units than when k is optimised. This is an approximate 95% confidence interval which runs from c.v. = 1.53-1.85 (*k* from 0.43 to 0.29). (b) The relationship between c.v. and sum of internal and external edge lengths (blue-gray and blue lines respectively), and inferred divergence times of human-gorilla (red line) and human chimp (dark green line using both human and chimp lineages, light green ignoring chimp lineage)(all measured as ratio of edge lengths from tips of these taxa to internal node of split vs length back to node where orangutan diverges).

Figure 5.9b shows the relationship between estimated relative divergence times, and also the total sum of edge lengths, as the rates across sites, which are assumed to follow a Γ distribution, become more extreme (all other parameters being simultaneously optimised). As expected, the external edge lengths get longer at an increasing rate, while the internal edge lengths increase more slowly, then apparently stop increasing and finally decrease. Divergence times initially decrease slightly, then, somewhat surprisingly begin to rise again. (this rise appears to be due to the internal edges, especially the common human-chimp-gorilla edge increasing in length more slowly). The human-chimp and human-gorilla divergence times then begin to converge, as the internal human-chimp edge begins to shrink in size relative to the external edges of human and chimp. Again there is an interplay of factors, which suggest we need to be reasonably confident of our model to know if our estimates are likely to be biased upwards or downwards. The reason for the stall and then decrease in the size of the internal edges, may be visualised as the more rapidly evolving sites (expected under as the c.v. rises) increasingly explaining the "informative" (continues next page)

changes supporting the internal edges as due to multiple hits. In the region of the optimal fit of k, divergence times are moderately stable. The difference in the light and dark green lines is the inferred divergence time of humans and chimps, varying depending upon whether the chimp lineage is excluded from this calculation on the suspicion it has slowed down (a possibility suggested by Horai *et al.* 1992, evident but not quite significant at the 95% level under our models by a clock constraint test, and not evident in the amino acid based studies of Adachi and Hasegawa 1995, where chimps if anything appear slightly faster evolving than humans).

The issue of standard errors on parameters in these models is an important one. A standard assumption is that the difference in G^2 statistics can be reliably used under a variety of circumstances. Asymptotically ($c \rightarrow \infty$) and under the null model (which includes knowing the true tree) both G^2 and differences of G^2 are expected to have a chi-squared distribution, with easily calculated degrees of freedom. Unfortunately, the very sparse nature of nucleotide data suggests that this is not true, and that G^2 and even differences in G^2 are usually compressed in their range and have lower expected values (see Reeves 1992, Goldman 1993, and chapter 6). There are two common ways of inferring the G^2 difference invoked by changing a parameters value (this change in G^2 then being used to construct a confidence interval, for example). The theoretically proper way is to alter the parameter of interest until the fit of model to data worsens by a specified amount, all the while reoptimising all other free parameters in the model. (Asymptotically and under the model, the difference in fit as one parameter randomly fluctuates about its optimal value is expected to be a chi-square random variable with one degree of freedom, which at $\alpha = 0.05$ implies that G^2 must change by 3.84 to be considered significant). An approximation used to avoid extra computation is to change the parameter of interest without reoptimisation (this approach is used, for example, in the program DNAML of Felsenstein 1993).

We evaluated the difference in estimated confidence intervals with and without simultaneous optimisation. Firstly without reoptimisation, the 95% confidence interval (C.I.) for the parameter k of the 1P gamma distributed model was estimated to be from 0.487 to 0.668 (with the optimal value being 0.568, for the 6 sequences of Horai *et al.* 1992). When we reoptimised other parameters, with k fixed to these boundary values, the difference in G^2 from the optimal model decreased, being 3.37 at the lower end and 3.43 at the upper end of the C.I. Using asymptotic approximations, this implied α rather than being 0.05 was really closer to 0.064 to 0.066, which is a relatively minor change. However, making a similar comparison on the best fitting invariant sites model with 3 parameters per edge (and 28 parameters in total), there is a more substantial change. The 95% C.I. of p_{inv} was 0.5733 to 0.6095 using the approximate method, but this was only an 84% C.I. when we reoptimised the other parameters (the difference in G^2 decreased by 2.05 at the lower p_{inv} value, and 1.89 at the other end of the approximate C.I. value). Thus the approximation appears to be getting worse the more free parameters there are in the model, and as this number increases it is expected that under asymptotic conditions this interval will be too narrow. A binomial C.I. on p_{inv} (which as already discussed has a binomial marginal distribution) is from 0.5777 to 0.6053, and Monte Carlo simulations by Goldman (1993b) suggest that such intervals are quite exact. This suggests that due to sparseness, both of the former two confidence

intervals are too wide, and it may be best to trust the binomial interval. Unfortunately, there is no binomial interval for the other parameters in the model (e.g. *k*, edge lengths etc.). It will be useful to know if estimating the distortion of the confidence interval for p_{inv} , by reference to the binomial, will yield a reliable calibration of the factor by which sparseness is altering the confidence interval of other parameters in the same model. Hopefully then such an estimate could also be transferred to similar models (similar in the mechanism of change, the tree, the total number of free parameters, and the overall fit) which do not have p_{inv} in them.

In either case, the confidence interval about p_{inv} is reasonably tight. This confirms that there are significant differences between the 1P vs all other models (the difference in p_{inv} being a change from about 0.6 to just 0.54), when it came to inferring the dispersion of rates across sites. A test of the difference in the means of two binomial variables is highly significant (the difference in means is > 0.05, with an expected standard error of only 0.010, so z is > 5, $P << 0.001$), and this is without taking account of the positive correlation of p_{inv} under the different models due to sharing the same data, which would increase the significance level.

5.3.9.1 Other results on this mtDNA data

A large number of these analyses were replicated using the minimum X^2 method of measuring fit of model to data. The minimum X^2 method sometimes outperforms ML methods on certain types of data (see Cressie and Read 1988). For this data, and these models, we found the performance of these methods very similar (data not shown). The size of the X^2 statistic to the G^2 statistic tended to be very volatile on the hominoid data, it would move from 60% of the size of the G^2 statistic to 60% larger than G^2 on moderately ill fitting models (with unoptimised parameters). X^2 tended to on average, appear larger than the G^2 statistic on the hominoid data. On the four rRNA sequences, it tended to be smaller, and generally more similar to G^2. This difference is most likely due to either the sparseness of the data (especially on cells with a count of 2, for example, when the expected value is less than 0.5) or equally due to an increased sensitivity to certain departures from the model (e.g. see Read and Cressie 1988). As an example of the sort of differences that using the X^2 criteria makes, the X^2 statistic at the G^2 minimum of the inverse Gaussian model of table 5.7 was 34.97 versus 37.78 for the G^2 statistic. When optimising explicitly by the minimum X^2 criteria, but keeping all other factors equal, the X^2 statistic dropped to 34.29. With this criterion the optimal shape parameter changed from 0.298 to 0.269, the over all tr/tv ratio from 2.69 to 2.70 (with $\Sigma tr: \Sigma tv1: \Sigma tv2 = 1.570: 0.353: 0.808$), with the sum of edge lengths increasing from 2.426 to 2.731 (this tends to reflect an increased emphasis on explaining some of the rarer patterns). That is, while the overall indication of goodness of fit often did not agree, especially on the worse fitting models, the estimates by the minimum X^2 criteria were, everywhere we looked, generally very similar to the ML estimates. Later in this chapter evaluations do suggest that the minimum X^2 method will sometimes have a slight edge over the ML method in specific situations. The overall impression was of very similar behaviour, with the most noticeable advantage of the X^2 method being that a quick perusal of the term $(\text{observed} - \text{expected})^2 / \text{expected}$ per cell offered a useful way to spot specific deviations from the model.

Recently we have had access to test versions of PAUP* (Swofford 1995), and have been able to further test our predictions of model behaviour. The model with which we will compare the earlier results is a 9 parameter time reversible model, with a numerical integration of the Γ distribution in 50 equal sized intervals, with each interval represented by the mean rate (a model like that used by Yang 1994). Optimising all free parameters (a total of 9 rates + 1shape + 9edges = 19), this model applied to the edited Horai *et al.* (1992) data had a log likelihood of -14,429.78 (versus -14,787.50 for the 3PΓ model), and estimated k = 0.247, with tr/tv = 13.12. The edge lengths of the homogeneous time reversible model (and the 3PΓ model in brackets for comparison) were: H: 0.0790 (0.0681), C .0225 (0.0206), P 0.0182 (0.0173), G 0.1044 (0.0834), O 0.2243 (0.149), S 0.2828 (0.169), CP 0.0316 (0.288), HCP 0..0306 (0.0217), and HCPG 0.0771 (0.542). The sum of all edge lengths was 0.793 (substitutions per site), while the estimated divergence times were CP 0.108 (with all African taxa used), HCP 0.369 or 0.445 (excluding chimps) and HCPG 0.565 or 0.603 (excluding chimps). The corresponding estimates from the 28 parameter 3PΓ model were k = 0.351, tr / tv = 7.63 sum of edge lengths 0.611, CP divergence 0.138 (with all African taxa used), HCP 0.427 or 0.502 (excluding chimps) and HCPG 0.601 or 0.638 (excluding chimps). Clearly this homogeneous general time reversible model fits the data much better (if we are not going to adjust for unequal base composition) and with fewer parameters than the 3P model, with a difference in G^2 (measured on the 4^t data) of 2 × (-14,441.07 minus -14,787) = 692.86 a large change (although not directly comparable to the changes seen under the order 4^{t-1} statistic, e.g. figure 5.7, since the likelihood there is compacted by about 2 times relative to the order 4^t likelihood). Such a change may be expected given the clearly unequal base compositions in this data (see section 1.9.2).

Other models taking into account base frequencies showed similar behaviour. The Hasegawa *et al.* (1985) (or HKY 85) mechanism has five distinct parameters, and is a submodel of the general time reversible model. If we look at this model plus Γ we have a likelihood of -14,446.7 and the parameters k = 0.232, and tr / tv = 11.77. The likelihood of this mechanism, plus p_{inv} was -14441.08, returning parameter estimates of p_{inv} = 0.626, tr / tv = 9.40, and a sum of edge lengths of 0.696. The estimated divergence times were CP 0.135 (with all African taxa used), HCP 0.398 or 0.466 (excluding chimps) and HCPG 0.576 or 0.610 (excluding chimps).

The shifts in parameter estimates in going to the general time reversible model from the K3P model, were partly predictable from earlier findings. The drop in k mimicked the jump seen in the difference of the Poisson models, to those allowing unequal transition to transversion rates (the same occurred with the p_{inv} HKY 85 model). The close relatives of the general time reversible model are showing the same trends seen under the K3P model, that is, k increases slightly as the model gets worse. The change in the tr/tv ratio was quite large, it is still fluctuating over quite a range, 13.2 - 9.4, amongst the models allowing for unequal base composition. The expectation is that still better fitting models (e.g. those better matching the distribution of rates across sites, probably easiest done by data editing into first positions, second positions etc., perhaps mixed with a 12 parameter mechanism of evolution) will see another increase in estimated tr/tv ratios. How high this ratio can go for this data is uncertain. In the

control region of humans, where fewer multiple changes may be hidden due to many intraspecific samples, the ratio seems very large (e.g. see Kocher and Wilson 1991, Penny *et al.* 1995).

The increase in the sum of edge lengths on the models allowing for unequal base composition was of some size, and to be expected with estimating increasingly unequal rates across sites (it is slightly larger than the 0.78 sum of edge lengths for the best K3P invariant sites model). The change in edge lengths was uneven, being most pronounced on the longer edges (e.g. the orangutan edge increased 51% while the increase for siamang was a large 68%, under the general time reversible Γ model). This accordingly decreased the divergence times of the more recent nodes. Again the Γ and the p_{inv} models differ in their predictions, with the p_{inv} models favouring older divergence times. Better fitting models may continue this trend, towards more recent divergence times (especially allowing for data editing). Overall, the changes in parameter estimates in going to the models which allowed for unequal base frequencies, were at least partly predictable. Many of them continued patterns seen earlier under the 3P model and submodels.

Interestingly, a preprint received from Dr Hasegawa just before completion of this thesis (Adachi and Hasegawa 1995) suggests that amino acid based ML models drop the estimated relative divergence times to as low as 0.14 for chimp-chimp, 0.28 for human chimp and 0.45 for human gorilla, all measured relative to the human orangutan divergence time. This raises a dilemma of its own. If these estimates are unbiased, then the human-chimp divergence time is only 3-4 million years ago if the favoured date of 13 myr ago is used for the human orangutan divergence. This makes the 16 million year old date favoured by Waddell and Penny (1995) look more likely, since there are now claims of substantial hominid finds of greater than 4 million years old (e.g. White *et al.* 1994). It also raises the question of how well the mtDNA dates are really agreeing with the nuclear dates (including those estimated by DNA hybridisation). It will be interesting to see if the four times greater coalescent time of nuclear genes is adequate to explain the discrepancy of the mtDNA date, or whether there may be other reasons. It is also interesting that the amino acids and Adachi and Hasegawa's (1995) models gave different relative dates to the 4-state model used here (even their simple Poisson model gives similar times).

How well the nucleotide model divergence dates match those of the amino acids is uncertain (the hope is they should match within sampling error when both sets of models are well tuned, else factors such as nonindependance of codon sites, and codon usage may be having an appreciable effect). The general time reversible Γ model and the amino acids (see below also) agree well on the human-gorilla date, but the human chimp date, and the chimp-chimp date don't agree closely, and are going in different directions. The first thing to do to check this will be to edit the data so as to better match just amino acid sites (removing tRNA and non-coding regions) and perhaps to split into coding positions to better match the distribution of rates across sites (followed by fitting mixed Γ, inv Gaussian and p_{inv} distributions to each subset of sites). Here, the biggest changes in estimated divergence times from the observed nucleotide mismatches appeared to come with use of amino acids, or else 4-state models allowing for unequal base

composition, and not just a distribution of rates across sites as anticipated in Waddell and Penny (1995). Clearly, these are all factors to bear in mind when making divergence time estimates, and their implications are considered further in chapter 6.

5.3.9.2 Concluding remarks to these ML single tree analyses

This concludes our main study of the behaviour of these models of evolution, and their application to real data. The results suggest that distribution of rates across sites parameters tend to become more extreme as the model becomes worse fitting (either due to the tree, the mechanism of change or parameters in the model not being optimised). This trend is evident with similar mechanisms of change, but substantial shifts in optimal values can occur when a crucial part of the model is deficient (e.g. not allowing for unequal tr / tv rates). Edge lengths tend to be hardest to define, when the exact distribution of rates across sites is unknown. Transition to transversion ratios do shift, but apparently not as much. These shifts are certainly enough to make one aware that accurate estimation is a matter of refining the model, as much as increasing the sequence length. The X^2 criterion performs similarly to the minimum G^2 or ML criterion. A possible trend with its use is that it places increased emphasis on the rarer changes (i.e. gives them more leverage). Accordingly it will tend to predict longer edge lengths (and a wider distribution of rates across sites) when there are more multiple hits than expected by the model.

There have been few studies to date of the systematic differences between models and the parameter values they return. Yang *et al.* (1994) considered some features of the interrelationship of Γ shape parameter and tr / tv ratios under the Hasegawa and Kishino mechanism of evolution. They did not find any particularly strong trends. From the studies made here, it would seem that their interest in the star tree as a neutral model on which to estimate parameters is most likely wrong. Better advice would appear to be: use the best tree found to make parameter estimates, and then compare these with other near optimal trees to gauge the likely magnitude of effect due to the tree. The fluctuations that come with different models are harder to ascertain, but if confidence intervals are of importance, an effort should be made to consider how much they change given different assumptions. One general way of doing this is used in Waddell and Penny (1995), but there is much room for refinement.

5.3.10 Approximate likelihood via approximations to Hadamard conjugations.

A potentially useful result in Szekely *et al.* (1993) is that the Hadamard conjugation (going from **s** to γ) can be expressed as a polynomial. Taking just the first term two terms offers a first order approximation which is easy, and computationally cheap, to calculate. Working with Dr Mike Steel we have derived the appropriate approximation for the inverse conjugation (γ to **s**) for the models discussed in chapter 2; that is for both the i.r. model, and also when there is a distribution of rates across sites. When starting with a vector description of a tree (as likelihood does) the result is a first order approximation of any sequence pattern $\mathbf{s}(T)_i$ for that tree, without needing to calculate any other sequence pattern probability. We will call these patterns q-$s_i(T)$,

where the q may be thought of as standing for quadratic approximation. For just those patterns i which are represented in a data set we calculate the quadratic-likelihood of a tree as,

$$q[\ln LR] = \sum_{i=0}^{2^{t-1}-1} \hat{f}_i \ln\left(\frac{\hat{f}_i}{q[s_i(T)] \times c}\right).$$

As with likelihood, an unobserved pattern contributes nothing to the likelihood so when $f_i = 0$ the term to the left of the summation takes on value 0. We expect the method to be most useful for calculating the approximate likelihood of large numbers of taxa, a situation where the exact Hadamard conjugation becomes computationally very expensive in terms of computer memory requirements, and increasingly expensive in terms of the total computations to recover the probabilities of a small set of patterns (often less than 400 distinct patterns even for 30 taxa)(a similar argument holds for other methods of calculating likelihood). This quadratic likelihood will also be a lot faster to calculate then the exact likelihood. It should be nearly as accurate as the exact likelihood for large sets of taxa which are closely related. This method can be modified (as a sum of terms) to accommodate a distribution of rates across sites, e.g. a Γ distribution. The main disadvantage of these approximations is they become poor as path lengths become large. Their main strength is expected to be in intraspecific work, or for species complexes. An exciting challenge will be to integrate this work with infinite sites ML estimates of population parameters based on coalescent models, and hopefully give rise to fast, but reliable, finite sites population genetics models. A full description of this method is in preparation in co-operation with Dr Mike Steel.

5.4 RETICULATE EVOLUTION IN PHYLOGENETICS

Here we describe some special cases of "tree selection" that were studied during the thesis. We firstly analyse what reticulate evolution means in terms of evolutionary trees and sequence patterns. In the second part of this section a special case of reticulate evolution, that of ancestral polymorphism, is considered. It is shown that Hadamard conjugations and ML methods offer a unique way to estimate the amount of ancestral polymorphism, and then via coalescent theory, infer the ancestral population size.

5.4.1 A likelihood model of reticulate evolution.

Here we describe a logical way of looking at reticulate evolution, especially when it is accompanied by sequence recombination within lineages descendant of the "hybrid" populations. By reticulate evolution we mean events such as hybridisation between species, where a lineage can trace its DNA to members of more than one "parental" species. If a DNA sequence is a distinct non-recombining locus, then its parentage has come from one or the other species, but not both. In a morphological or behavioural data sets, with their multiple genetic loci, it is often observed that the descendant population is polymorphic for its parent species traits, followed by sorting due to loss of parental traits (e.g. Waddell 1990, p26-30). Some consideration has been given to the implications for parsimony analysis (McDade 1990, 1992), but to date little work has been done to describe coherent models and allow the development of likelihood models.

At the sequence level a good analogy to the situation that occurs with independent morphological characters, is that a stretch of sequence is inherited from two distinct ancestral lineages and that this locus then undergoes recombination along its length. Consequently, following further evolution, we cannot be sure of where the boundaries of pieces of DNA with separate histories lie. Here we show that recombination within a DNA sequence makes specific predictions, which should allow more reliable assertion of the probability that a sequence underwent reticulate evolution.

Figure 5.10 shows a reticulate evolutionary event, with edge lengths proportional to time. We have started with a slightly complicated situation to show how easily it sorts itself out into a sum of binary trees. Two distinct reticulate events have occurred to give species b and c. Four separate ancestral populations have split off from the lineages leading to extant populations, and persisted for variable periods of time before one pair, then the other, form a hybrid populations with all the "parents" going extinct as independent lineages. Associated with each reticulate event is a proportion indicating how much genetic material from each species becomes fixed in the hybrid population.

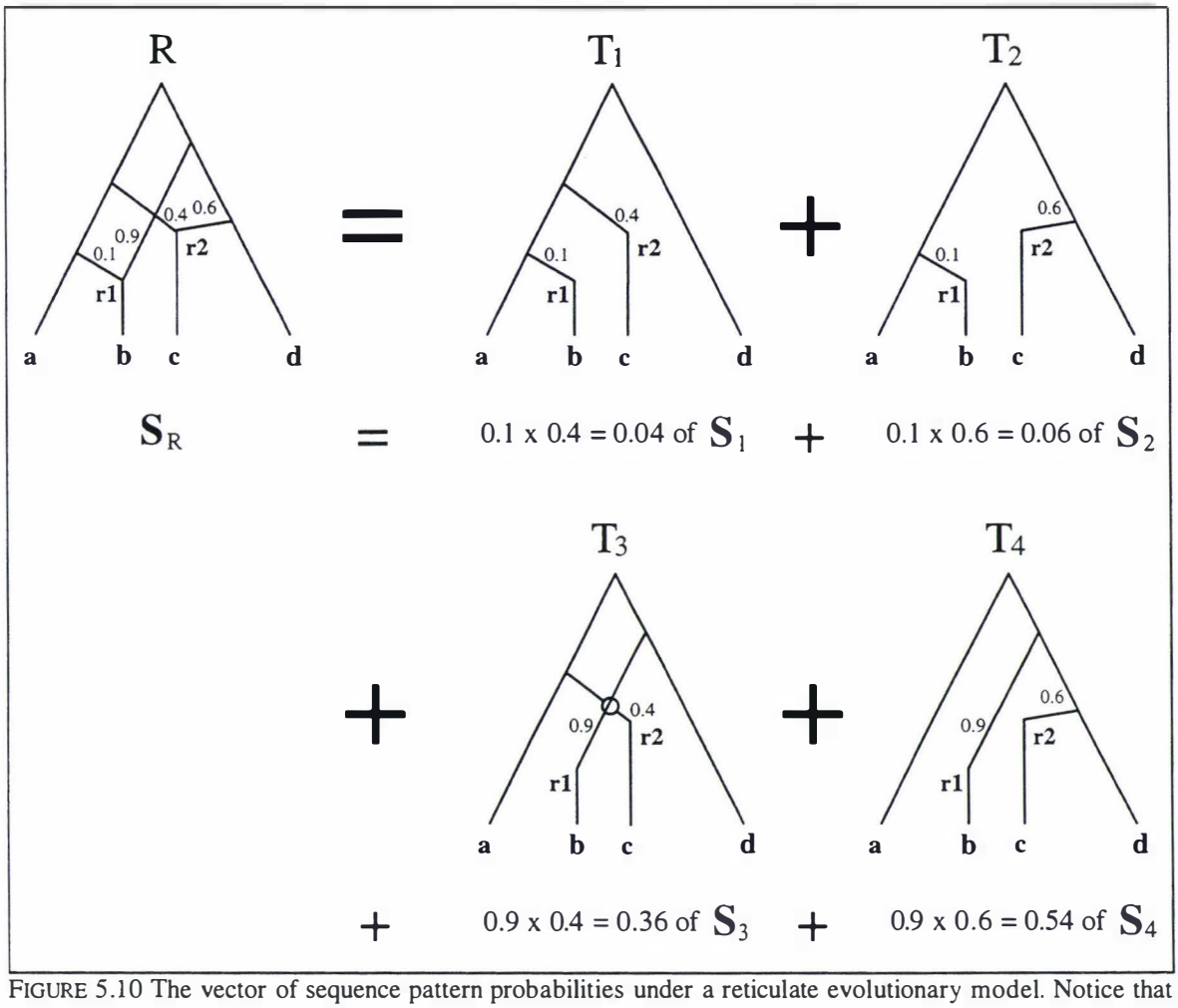


FIGURE 5.10 The vector of sequence pattern probabilities under a reticulate evolutionary model. Notice that a reticulate phylogeny can be broken up into a set of weighted binary trees. The proportion of sites from each parent fixed in the "progeny" (in this case a new lineage) is shown in the reticulate phylogeny as the number associated with each "parental" edge. Given assumptions of independent evolution of sites we can predict the proportion of sites in the extant species that have each evolutionary history. Consequently we can calculate the probabilities (i.e. the likelihood) of sequence patterns under the reticulate phylogeny, and it is then a numerical step to find the ML solution.

If we track a sequence site moving through the reticulate phylogeny, it could have had one of four different histories, each represented by a weighted binary tree. Indeed, considering the situation generally, if there are x reticulate events in a phylogeny, then at each event a site is involved in a binary split the site follows one or the other path to the present. Each of these events is equivalent to a split in a weighted tree, so the total number of possible histories for a site becomes 2^x, i.e. 2^x weighted trees describe the set of possible histories a site may have had on the reticulate phylogeny with x hybridisation events. Given the assumption that which way a site goes at a reticulate event is a random variable with stated probabilities, and assuming that evolution going down the tree is independent in each lineage, then by simply multiplying the probabilities at each reticulate event determines the probability a site followed a particular history (weighted tree, again see figure 5.10).

We generate the trees inferred by a reticulate phylogeny by labeling an ancestor going most to the left as 0, and the one going most to the right as 1. We can then use a binary number to

describe the pattern of forks, and the base 10 equivalent of this number is the trees index minus one. For example arranging the reticulate events from left to right, then the first tree is going to be 00 and the index equals 00 + 1 = 1. The next 3 trees will be 01, 10, and 11, with indexes respectively of 2, 3 and 4 (as shown in figure 5.10).

If a particular site incurred no more than one substitution in its history, then this site would either be uninformative or else show a pattern conforming to one of the edges in the weighted tree describing its history. However, under all Markov mechanisms of substitution it is expected there will be a spectrum of possibilities for multiple changes. Given each weighted tree, it is possible to calculate what these are. These vectors of sequence pattern probabilities are **s** vectors, and as figure 5.10 shows a unique vector is associated with each weighted tree. If we multiply each entry in these **s** vectors by the probability that a site followed a certain weighted tree, then sum them together, we obtain the probability of all sequence site patterns evolved on the reticulate phylogeny, according to that history (weighted tree). Assuming all sites to be unlinked (i.e. the probability that a site has one or other parent is statistically independent of the parentage of any other site) and to be evolving by the same mechanism, then we can take $\mathbf{s}_r$ (a sum of weighted tree model site pattern probabilities, with the *r* standing for reticulate) to describe the site pattern probabilities for an sequence of any length evolving in accordance with the reticulate phylogeny. In real sequences, recombination (and independent chromosome segregation for sites on different chromosomes) is the process by which sites become unlinked. By an argument like that presented in chapter 2 (e.g. appendix 2.2), even if there is some degree of local linkage then as long as this source of correlation falls of at a certain rate, the model is guaranteed to be consistent if all the other assumptions are met.

Supposing the sequences have multiple hits, then there could well be many misleading substitutions to obscure the history of sites that had changed just once. With multiple changes it may be difficult or impossible to assign sites to separate histories. In such cases we can infer the reticulate history of these sequences, by using a maximum likelihood model that does not require sites to be separated according to their history. The aim is to find the set weighted trees in specified proportions (consistent with a single reticulate phylogeny) such that the sum of their predicted sequence patterns best fits the observed sequence data (by the G^2 criterion). Notice that while there are 2^x (where x is the number of reticulate events) trees involved in these calculations (with 2t-3 times this number of edge weights), all of their parameters (edge indices, and edge lengths) and their relevant contribution to the overall model are predicted by the principal parameters of the model, the weighted reticulate network or phylogeny. There are just $2t-3 + 3x$ edge weights in the reticulate phylogeny, plus another $2x$ parameters to determine the probabilities of a site following either descendant edge at a reticulate event (where t is the number of taxa, and x is the number of reticulate events). Because the number of underlying model parameters is small compared to the number of edge weights in the trees describing a sites possible histories, it is obvious these edge weights are highly constrained and non-independent. These restrictions between the edge weights and patterns in the weighted trees in turn allow predictions to be made of what we should expect to see if reticulate evolution is the explanation.

An interesting implication of this representation of a reticulate phylogeny, is that even if the phylogeny is estimated accurately, we cannot estimate the time of events r_1 or r_2 exactly, even if a molecular clock holds (see figure 5.10). The best we can do in the case of r_1 is to use the calibrated height of the node on T1 where **a** and **b** have their last common ancestor as an upper bound on the time of hybridisation (in the case of r_2 the bound is the time of the last common ancestor of sequences **c** and **d** on tree T2 or T4). The reason for this is straightforward; we have no other specific information on how long the ancestors of the hybrids may have been separated from any other sampled lineage. Tightening this bound requires further sequences to intersect the edges leading to the founding of the ancestral populations, while edges intersecting hybrid lineages help establish a lower bound on their formation. This finding is important and it must also apply to the type of reticulate model used in Bowcock *et al.* (1991), thus their suggestion that they can maximise the likelihood of the time of mixing in a reticulate phylogeny based on alleles must also be incorrect. Instead, they are estimating the time that the hybrids parental populations last had a common ancestor with other populations in the study (with the additional, perhaps unrealistic assumption that this time was identical in both cases). Only by the further, dubious, assumption that these two parental populations immediately hybridized can you claim to be estimating the time of hybrid population formation. Note that the model described here should be quite applicable to population genetics, where recombinations are expected to occur between markers (or alleles) scattered across the nuclear genome (e.g. micro satellites).

If there were no parallel changes, we could confidently reconstruct the reticulate phylogeny. This requires removal of parallel and convergent changes. In this respect, a Hadamard conjugation of the data could be useful prior to such an analysis. The Hadamard conjugation is not exactly correct for this type of model because it assumes the data was generated by one tree. Here instead, the data are generated by a set of different trees and mixed up at the sequence pattern level, and not the gamma level. To make exact corrections in going from **r** to ρ to give $\Sigma\gamma(T)$, requires a sum of logarithms of r_i for the different trees. This is not the same as the log of a sum of r_i, which is what happens if sites are not separated according to history (i.r. sr is a Σ**s**(T)i. However, we expect it can still be effective in down weighting patterns more likely to be due to multiple hits rather than an edge on one of the potentially many trees. An example of this is given below, but a more detailed examination of how effectively this is done while changing the parameters of the model would be useful.

5.4.2 ML methods to estimate degrees of ancestral polymorphism.

Coalescent theory tells us that when there is a closely spaced divergence of species, there is a finite and sometimes quite high probability of a DNA locus (without recombination) having evolved by a different tree to the species tree (e.g. Nei 1987, Hudson 1990, 1992). A good example of this is the divergence of the African apes and the hominid lineage. It appears fairly certain that human and chimp are closest relatives (see Waddell and Penny 1995 for a summary of the data). However, if the time of branching of the gorilla was not so long (say 1 to 3 million years) before this, then there is the possibility of some sequences having evolved according to a tree linking human and gorilla sequences as closest relatives. There is also an equivalent

probability of a chimp and gorilla sequence being closest relatives (see Hudson 1990 for a review of this type of process). If we have a region with recombination in it, then there is a real possibility that part of the sequence could have evolved on one tree and part on another different tree. Given a long stretch of DNA, with lots of recombination spots on it, a reasonably high rate of recombination, and a moderately large ancestral population size, we would expect a mixture of sites with different histories. In a region with high rates of recombination evenly dispersed along the sequence, then potentially every segregating site (from the ancestral population) can be treated as an independent locus (for example in the formulae of Hudson 1992).

5.4.2.1 Examining the human-chimp-gorilla divergence

There is a region of sequence which, during the studies for Waddell and Penny (1995), was not fitting ML analyses as well as expected. Following examination of the cell by cell X^2 fit of this region the problem appeared to be an excess of sites grouping human-gorilla (HG) and chimp-gorilla (CG), whereas most other patterns fitted the model homogeneous Kimura 3P clock model reasonably well. The sequences of interest are non-coding nuclear DNA, from the primate β-globin gene cluster, a thorough review of which is given in Bailey *et al.* (1992). Heightening suspicion that this was not a fault of the model, but rather a feature of molecular evolution, was the mtDNA sequence of Horai *et al.* (1992) not showing similar anomalies under the ML URAS models, despite these sequences clearly being much more divergent.

The sequence that will be used to illustrate this new method of analysis is a 10 kilobase (kb) stretch from the β-globin region, known as the psi-eta (ψη) region (Bailey *et al.* 1992). The alignment used consists of the species human, common chimp, gorilla, orangutan, and rhesus monkey, with all deletions removed (Waddell and Penny 1995). The findings from this region do not differ qualitatively from our analyses of the 9kb of γ (Bailey *et al.* 1992), or 6kb of δ--β intergenic region which are downstream. Further, it is known that recombination occurs at an appreciable rate in the β-globin locus (Kazazian *et al.* 1983, Chakravarti *et al.* 1984), and is detectable amongst modern human populations despite evidence for their surprisingly recent origin from a small population (e.g. Cann *et al.* 1987, Vigilant *et al.* 1990, Rogers and Harpending 1992). If we plot out the sites grouping HC, HG and CG in these regions (see figure 5.10b), there is no observed clustering as would be expected if only one or two recombinations had occurred in these regions (and there were no convergences or parallelisms to give these patterns). Overall there are three transitions and transversions supporting HC, 2 transitions and one transversion supporting HG, and two transitions supporting CG. The lack of parsimony informative sites from position 512 to 6029 appears to be a random feature, as there are many other site substitutions in this region.

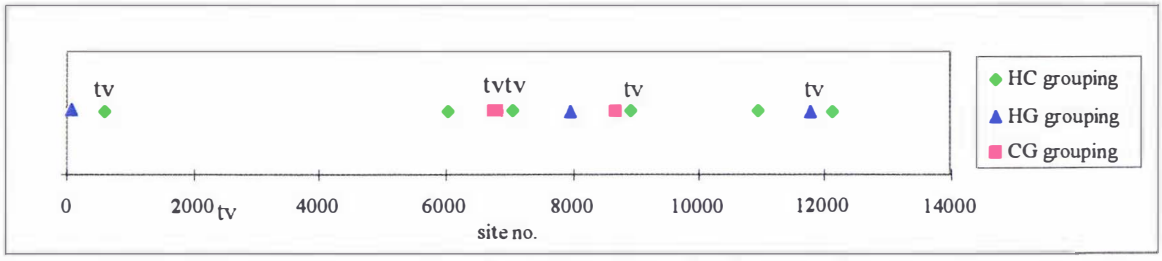


FIGURE 5.10a The position of parsimony informative sites along the ψη region (tv marks transversions)(the red square at 6700 represents two closely spaced sites). Our data set is slightly different to that of Bailey *et al.* (1992), due to a deletion in one of the five taxa, it does not include an additional CG transversion. (Figure 5.10b next page).

In the present day, we cannot estimate the frequency of each of these site histories directly because there are potentially many convergent and parallel substitutions. If these can be removed, we would once again be in the business of estimating the human-chimp-gorilla ancestral population size with the coalescent formulae given in Hudson (1992).

For this study, the most appropriate formula with which to estimate ancestral population size is conditional on the probability of observing each type of segregating site. Given the standard assumptions of a population in equilibrium, Hudson (1992, equation 9) gives this probability as,

$$\text{P(site congruent to species tree)} = 1 - \text{P(non-species tree)} \approx 1 - \frac{2e^{-T}}{3}\left\{\frac{1}{T+e^{-T}}\right\} \qquad (5.4.2\text{-}1)$$

where T, the sole variable, is measured in units of 2N generations (N being the diploid population size). T can, in turn, be expressed as $t_{div}/(t_{gen}2N)$ where t_{div} is the time between divergence of gorilla and humans and chimps, t_{gen} is the average effective generation time of ancestral hominoids at this time, while N will be the effective population size of these ancestors. In section 5.4.2.3 we show that we have reasonable estimates of these other parameters and so can solve for our ancestral population size, once we obtain good estimates of P(non-species tree).

Throughout this section we assume a constant effective ancestral population size in the period prior to the divergence of the human-chimp ancestor. We do this because most coalescent theory is based upon this assumption, and it seems the most reasonable starting point given recent studies that modern apes appear to conform reasonably well to these expectations (e.g. Morin *et al.* 1994). This method can, as appropriate, incorporate improvements to general coalescent theory.

5.4.2.2 Estimating ancestral diversity free of the effect of multiple hits

We have developed three main ways of removing the parallel and convergent substitutions, in order to estimate the ancestral population size free of the effects of multiple hits. One way is via a Hadamard conjugation, another is a new ML model (plus close approximations), while the third is a modification of the more traditional ML tree estimation of Felsenstein (1981a). In either case, the essential feature is to estimate a reliable probability for convergent and parallel (continues next page)

substitutions which is not distorted by the potential presence of excess human-gorilla and chimp-gorilla patterns (due to the mixed histories of sites).

The first technique begins by applying the 4-state Hadamard conjugation to the data. The next step is to take the sum of the two non-species tree signals (HC and HG), divided by the sum of these two values plus the signal for the species tree (HC patterns). This is our first estimate of the probability of a site following a history other than that of the species tree following corrections for multiple substitutions. In the case of the ψη data, and after correction under the general Kimura 3P i.r model, this gives HC (2.71 tr, 3.3 tv1, -0.02 tv2), HG (1.70 tr, -0.02 tv1, 1.08 tv2), CG (1.73 tr, -0.07 tv1, -0.02 tv2), so our estimate of P(non-species tree) is (HG + CG) / (HG + CG + HC) = 0.424. Later, it is shown that ML optimisation assuming a Γ distribution estimates k = 3.51 (although the difference in fit from the i.r. model is not significant). Correcting the data with a Hadamard conjugation, assuming a Γ distribution with k = 3.51 gives P(non-species tree) = 0.413, a small change.

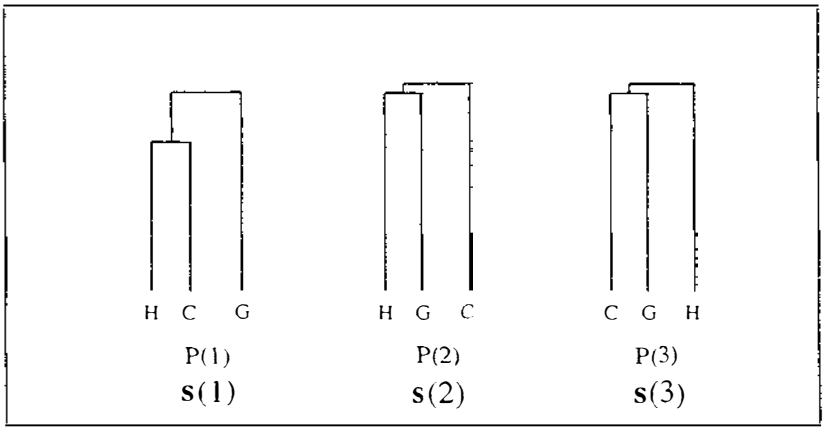


FIGURE 5.10b The three weighted trees used to represent the different histories that independent loci can show if the HC edge of the main tree (the left most tree) which follows the population history, was not sufficiently long to ensure the loss of all ancestral allelic diversity. The time of the splitting of the original population is indicated by the horizontal dash-dot-dot line. The dotted lines at the top of the trees indicate the edge leading to the remaining taxa in the analysis. The effective size the population prior to its division will leave two distinct features: One will be the proportion of sites having each history (represented by P(1) to P(3)), while the other will be the length of the edges (or the variance of the expected allelic divergence time) in the different histories (our diagram shows just expected or average edge lengths). Both of these features can be used to infer information about the size of the ancestral population. Given a mixture of sites with these three distinct histories, we can apply ML methods (as described in section 5.4.1) to estimate the proportion of sites following each history and the average edge lengths associated with each history. If we assume a molecular clock we can restrict the model further by making trees two and three have identical edge weights. Another restriction may also be reasonable is that the ultimate coalescent time of all three alleles be non-independent (dependent upon the size of the edges in the species tree and the ancestral population size), and possibly even expected to be equal (we will need to check this conjecture with the expected coalescent times of alleles with the different histories).

Our second technique is based on the ML model described in the section 5.4.1. If the only part of the tree showing close spaced branching is about a single trichotomy, then there are just three distinct trees in the reticulate model. The next step is to optimise the likelihood of this

model; in this case we are confident that the species tree is (((H,C),G),O) and two subtrees are (((H,G),C),O) and (((C,G),H),O) (where O stands for all outgroups to the group of three). All three trees will have very similar external edge lengths (see figure 5.10b for a fuller description). After optimising the likelihood of this model, the ratio of interest is the sum of the internal edge lengths on the two non-species trees, divided by the sum of these two terms plus the edge length on the species tree. This approach should be more exact than the Hadamard conjugation method for two reasons. It should not be inflating the errors in the way that the direct non-linear transformation does (see chapter 4). And this reticulate phylogeny ML method model does not assume the log of a sum is a close approximation to the sum of logs, though this approximation should be fairly accurate with low rates of change. The accuracy of the Hadamard conjugation method may be assisted by the three external edges, H, C and G being nearly equal (this requires further exploration).

At relatively low rates, a close approximation to the three tree ML model can be made by selecting not only tree entries in the $\gamma(T)$ vector, but also simultaneously optimising the two sets of entries corresponding to the internal edges (counting transitions and transversions) of the two non-species trees. In the analyses for Waddell and Penny (1995) it was found that a Kimura 2P fixed tr/tv ratio clock ML model gave a reasonable fit for our purpose here. We have also allowed for a Γ distribution of rates across sites, in case there was any detectable effect of non-i.r. rates. The fit of the best single ML tree to this data (according to these parameters) was the (((H,C), G), O), R tree, which had a G^2 of 112.38 (with k = 3.52, but not significant since when the model was set to i.r. G^2 dropped less than 2 units to 114.22)(the tr/tv ratio was optimized at 2.31). Following reoptimisation, allowing for the two non-species tree internal edges in the reticulate model (and fixing them to be of the expected equal size, so adding just one parameter to the model) the G^2 fell very significantly to 100.13 (by a χ^2 distribution with 1 d.f. $P << 0.001$). Interestingly, the Γ distribution shape parameter k fell to just 15.04 (while the difference from the i.r. model was then just 0.01 G^2 units); this finding is concordant with our earlier claim (section 5.3.7) that in closely related models shape parameters tend to become more extreme to compensate for inadequacies in the worse fitting model. Taking the estimated internal edge weights in this ancestral polymorphism ML model, and using them as our estimates of the number of sites following the species tree or not, gives P(non-species tree) = $2 \times 2.81 / (2 \times 2.81 + 6.42) = 0.467$ (a slight increase over the previous estimates made with the Hadamard conjugation).

The third method of removing the effect of multiple hits is an approximate ML method. The idea is to implement ML optimiastion on the species tree, but not allow the site patterns suggesting a HG or CG derived state to influence the likelihood at all (i.e. effectively removing them completely from the analysis). The internal edge weight of this tree gives the corrected number of sites showing the derived pattern HC. Then using edge weights of this tree the probability of sites showing HG or CG is calculated. These last two numbers need only be equal if we assume a molecular clock, which is appropriate in this case (we will use these numbers as correction factors). Our model fitted much better than the previous standard single tree model

(G^2 was 102.01 a drop of $10.37G^2$, while k was 7.21 which was only 0.37 G^2 units from the i.r. model, and tr / tv was estimated at 2.28). The internal (HC) edge weight of this tree was 0.00064 × 9892 = 6.35, while our model estimated the frequency of a HG (or CG) transition bipartition pattern was 0.45, while a transversion bipartition pattern had an expected frequency of 0.04 (these numbers being $f_i = cs(T)_i$). Since we observed 5 sites with HG or CG bipartition patterns and just 0.99 was expected, this leaves 4.01 substitutions attributable to ancestral polymorphism. Thus the estimate of P(non-species tree) = 4.01 / (4.01 + 6.35) = 0.387. This is a slightly lower estimate than previously. This may be due to this method inferring the total number of substitutions expected on the HC edge (including those that will not show up as HC observed bipartitions due to other substitutions at the same site), but not adding this amount to the observed HG and CG bipartition site patterns.

An advantage of this approximation, is that it can easily be used with available ML one-tree programs such as DNAML (Felsenstein 1993). To do this firstly, remove the all HG and CG patterns from the data (so they do not contribute to the likelihood), and estimate the weighted species tree. Save this, and then ask the program to evaluate the likelihood of each HG and CG bipartition pattern given the weighted tree (there a 12, i.e. AACCC, ... , TTGGG, of these for HG and 12 more for CG). The sum of these pattern probabilities, multiplied by the sequence length, yields the required correction factors.

It is straightforward to explain why this last modification works better than the standard ML tree selection model. If there are sites with different histories due to ancestral polymorphism, then unless these sites are separated, the reconstructed tree need not accurately reflect the (weighted) species tree. What happens is that the informative sites from the other trees are interpreted as convergences or parallelisms by this "one tree" model. These rare patterns can have surprising leverage on the overall likelihood (or G^2 goodness-of-fit), and consequently the external edges (i.e. for H, C, G) tend to noticeably (≈ 5 to 20%) increase in length to make these patterns more likely under the model. At the same time the internal HC edge will shrink, since some of the HC patterns are inferred to be the results of multiple hits due to the increased external edge lengths. By making the modification of removing the HG and CG bipartition patterns, they can no longer exert this bias on the one-tree model.

If we make estimates under the one tree model, the result is: the estimated length of the internal HC edge is 6.23, the inferred number of multiple hit bipartition transition patterns HG or CG is 0.59, while transversion patterns total 0.05. As above, the corrected number of bipartition HG and CG patterns becomes 5 - 2 × 0.59 - 2 × 0.05 = 3.72. The overall estimate of P(non-species tree) is (3.72)/(3.72 + 6.23) = 0.374. Clearly there are two biases operating in opposite directions to keep the final ratio near previous estimates. The first is the underestimate of HC patterns, the second being an underestimate of the genuine HG and CG patterns (which is more severe than in previous cases due to the longer external edges). The expectation is that the leverage of leaving in the non-tree patterns in the one-tree model will decrease the support for genuine HC, HG and CG substitutions about equally. Thus if this downward bias is x, then (HG + CG - 2x) / (HC + HG + CG + 3x) will have a downward bias as, by definition, HG + CG are

unlikely to be greater than 2/3(HC + HG + CG) (and the smaller their sum is, the more severe the downward bias). This bias will result in an underestimate of the ancestral population size.

The overwhelming conclusion from all these different methods, is that there very few HC and HG bipartition patterns expected to be due to multiple hits (in fact only about one of the five observed). Later we will show that these differences in observed and expected frequency are statistically significant, and this difference does not occur when counting up the convergent patterns in other parts of the tree. In the discussion of this section we present evidence that this is not a unique feature of this gene sequence or even of the β–globin cluster, but is suggested by other nuclear sequences. Next though, we make our estimates of our ancestral human-chimp-gorilla population size.

5.4.2.3 Solving for ancestral population size

The computations of the previous section bring us to the point of solving equation 5.4.2-1 for T, then from T estimating the ancestral population size given an estimate of the time interval between human-chimp and gorilla divergences, and African hominoid generation times. Our estimates of P(non-species tree) were 0.413, 0.467, 0.388 and 0.374 for the four methods (with the second method expected to be the best estimator, which will also give the largest estimate of population size). The estimates of P(congruent with the species tree) are just 1 minus these numbers, equals 0.587, 0.533, 0.612, and 0.626, respectively. Solving equation 5.4.2-1 for T with numerical methods, yields T = 0.408, 0.313, 0.456, and 0.483 respectively. These numbers are slightly more spread out due to the non-linear solution, but still in reasonably good agreement.

Lastly we require estimates of the duration of the HC lineage, and also the generation time of our ancestors in order to solve for the ancestral diploid population size $N = t_{div}/2t_{gen}T$. Ideally we want an estimate of the difference in divergence times of human-chimp from gorilla which is least affected by ancestral polymorphism (there are no fossils to help make this estimate directly). Because the effective population size of mtDNA is only 1/4 that of the nuclear genes, it should offer a useful estimate, which, as calculated earlier in figure 5.8, is between about 0.13 and 0.16 of the total time back to orangutans. Assuming that the African hominoid-orangutan divergence was about 16myr ago (see Waddell and Penny 1995 for a discussion of this point) the HC lineage existed for about 2.1 to 2.5 million years (myr)(i.e. 0.13 × 16 myr to 0.16 × 16 myr). Another source of data which can estimate this difference in divergence times is DNA hybridisation. Estimates from DNA hybridisation distances are potentially the most reliable, since under the assumption of a constant ancestral population size, the expected internal edge length of a tree built from them is exactly the species tree edge lengths plus the added height of the mean coalescent time. This last term disappears if we are comparing two nodes (e.g. HC vs HCG) where population size remains constant (Dr R. Hudson, pers. comm., provided a proof of this conjecture).

Unfortunately, in actuality, there are substantial differences in the estimated duration of the HC lineage from DNA hybridization data, probably due to experimental errors in all studies (Dr R. Britten, pers comm.). The size of the internal edge, relative to the total time back to the

human-orangutan divergence, is in the range of 0.04 to 0.14 for the DNA hybridisation data of Sibley *et al.* (1990) and about 0.26 for the DNA hybridisation data of Caccone and Powell (1989). We will take a value of 0.2 as an approximate average, which multiplied by 16myr gives 3.2 myr as the difference in divergence times of human, chimp and gorilla. Given the importance of this estimate for understanding human evolution, it would be highly desirable to have the DNA hybridisation experiments redone to the accepted protocols (Britten *et al.* 1990), and with an experimental design that allowed accurate measurement of the different sources of error (Felsenstein 1987).

Lastly, we require an estimate of the generation time of our ancestors. Human's have a generation time of about 18 to 25 years; a figure which doesn't fluctuate as much as one might suspect because the age of first reproduction (governed by the onset of puberty) and the age of last reproduction both tend to be negatively correlated, and both are correlated to the quality of nutrition. The average generation time of chimps is less. The age of first reproduction for females is as low as 10 years, but again females may remain fertile until they are in their mid thirties (e.g Goodall 1990). Males tend to reproduce at somewhat older ages as stamina, body size, and social ranking tend to be important attributes for reproductive success (Goodall 1990). An average generation time of 15 years seems reasonable. For gorillas, females tend to become fertile slightly (a year or two) younger than chimps, but if anything, males need on average to be older to have reproductive success, so again a generation time of 15 years seems reasonable. Given that all living African hominoids have a generation time of this order, and given a similar generation time for orangutans, it seems likely our last ancestors with chimps and gorilla had a generation time close to this (and a range of 12 to 18 years would almost certainly cover it).

So now taking our estimate of T = 0.313, t_{div} (from mtDNA) as 2.3 myr, t_{gen} as 15 years we obtain N = 2.3myr/(0.313 × 15 years) = 490,000 diploid loci, or 245,000 individuals. Such a population size is similar to that estimated for either chimps or gorillas as a species (e.g. Morin *et al.* 1994, Ruvolo *et al.* 1994). In contrast the effective population size of humans (*Homo sapiens sapiens*) is put at between 4,000 and 40,000 individuals for the last 500,000 years (e.g. Rogers and Harpending 1992, Takahata 1993). We suspect that this estimate is on the high side because our analyses of other parts of the β globin cluster (e.g. the γ globin region) are giving estimates of P(congruent with species tree) of closer to 0.75. This gives T = 0.77, so that the effective population size is then estimated at 100,000 individuals which is close to the population size of a subspecies of either chimps or gorillas.

While an effective population size of 245,000 seems very large given today's subspecies (which appear to have been reproductively isolated from each other for something of the order of a million years, Morrin *et al.* 1994), it may be reasonable. This is because the last 2 million years have been marked by climatic change, with periodic reductions in habitable range for all the great ape species due to arid ice-age conditions reducing forest cover in both Africa and Asia. At present we are making an alignment of all the nuclear sequence available for these taxa (with suitable outgroups) in order to get a more exact estimate of P(incongruent). Alternatively our estimate of population size may be biased upwards by assuming too old a date for the orangutan

lineage, or too long a duration of the HC lineage, as well the preliminary results of other genes putting T closer to 0.7.

5.4.2.4 Testing the adequacy of this model and our conclusions

We require tests of two hypotheses: firstly that there are significantly more sites supporting HG and CG than the model predicts. secondly, the remainder of the model fits with the expectation and does not underestimate the number of parallelisms and convergences in the data. We propose to test the first feature with a one tailed binomial test that the observed frequency of HC and HG patterns is greater than expected under the model. The second feature will be tested with a one tailed binomial test that the observed number of other parallelisms and convergences (bipartitions of HO, HR, CO, CR, GO and GR) are not significantly greater than expected under the model. This second feature is in fact a stringent test of the adequacy of the model, since there are substantially more parallelisms and convergences expected involving the long orangutan and rhesus monkey external edges than there are involving the much shorter human, chimp and gorilla external edges. Table 5.8 shows the observed and expected frequency of all the bipartition patterns for this data and the associated models. Table 5.8 shows that the only patterns which are present in excess of the expected are the HC and HG patterns. The only reason that the HO, ..., GR patterns approach significance is a very high count of GR bipartition transitions. Some of these could possibly be due to CpG dinucleotide sites which induce C $\rightarrow$ T or G $\rightarrow$ A substitutions at a much increased rate. The reason for this is that the modified cytosine base Cp (5-Methylcytosine or 5-mC) tends to deaminate to T (thymidine) resulting in the error correction enzymes being unable to spot the error (e.g. Perrin-Pecontal *et al.* 1992).

(table next)

Table 5.8 Observed and expected numbers of patterns for 10 kb from the ψη region. On the left hand side are transition bipartitions, while on the right hand side are transversion bipartitions. The expected frequencies are those estimated by ML for a standard one tree Γ homogeneous K2 ST clock model (ML St.), this model modified by ignoring HG and CG patterns in optimisations (ML Mod.), and the approximation to the three trees ML model discussed in the text (ML Anst.). The X^2 statistic is calculated for each cell of the ML Anst. model only. H: human, C: chimp, G: gorilla, O: orangutan, and R: Rhesus monkey.

	tr	ML St.	ML Mod.	ML Anst.		tv	ML St.	ML Mod.	ML Anst.	
Pattern	Obs.	Expt.	Expt.	Expt.	X^2	obs	Expt.	Expt.	Expt.	X^2
H	38	41.96	41.75	41.25	0.3	17	18.20	18.4	17.93	0.0
C	42	41.96	41.75	41.25	0.0	22	18.20	18.4	17.93	0.9
HC	3	4.43	4.45	4.51	0.5	3	1.71	1.8	1.79	0.8
G	52	45.80	45.75	45.30	1.0	9	19.87	20.17	19.70	5.8
HG	2	0.59	0.45	0.48	4.9	1	0.05	0.04	0.04	21.0
CG	2	0.59	0.45	0.48	4.9	0	0.05	0.04	0.04	0.0
HCG	51	51.90	51.40	51.30	0.0	23	20.99	21.42	21.02	0.2
O	93	95.38	95.32	94.67	0.0	44	40.89	41.62	40.75	0.3
HO	0	0.59	0.45	0.48	0.5	0	0.05	0.04	0.04	0.0
CO	1	0.59	0.45	0.48	0.6	0	0.05	0.04	0.04	0.0
GR	8	2.35	1.85	1.95	18.8	1	0.23	0.18	0.19	3.4
GO	1	0.84	0.65	0.69	0.1	0	0.08	0.06	0.06	0.1
CR	0	2.11	1.65	1.73	1.7	0	0.21	0.16	0.17	0.2
HR	1	2.11	1.65	1.73	0.3	0	0.21	0.16	0.17	0.2
S	348	348.93	353.71	352.16	0.0	164	157.60	160.98	158.39	0.2

Table 5.9 Testing observed and expected frequencies of bipartitions for 10 kb from the ψη region. After ML optimiastion on all patterns we have grouped bipartition cells into 4 main types for testing at the 0.05 level of significance (0.025 level with Bonferroni correction)(all tests made are one tailed exact binomial). The only patterns present in a significant excess of the expected numbers are the HG and HC patterns. The HO, .., GR patterns approach significance only because of a very high frequency of GR transition patterns; a trend shown by no other pattern (see table 5.8).

Type of BP pattern	ML St.	ML Mod.	ML Anst.	ML St.	ML Mod.	ML Anst.
	Obs.	Expt.	Expt.	Expt.	$P(X \geq obs)$		
singletons	829	828.79	837.86	829.32	0.497	0.626	0.505
internal edges	80	79.03	79.07	78.63	0.456	0.458	0.438
HG, CG	5	1.29	0.99	1.04	0.010	0.003	0.004
HO, HR, CO, CR, GO, GR	12	9.42	7.35	7.73	0.239	0.071	0.093

5.4.2.5 What this population size estimate may be telling us about human evolution

There are other ways of inferring ancestral polymorphism also. One could measure the fluctuation in inferred divergence times of many genes; building up a distribution of many of these would give a clue as to the extent of ancestral diversity (this approach is mentioned in Waddell and Penny 1995). A major problem is that unless a loci is very long, the stochastic variability of estimating divergence times (see Waddell and Penny 1995, and chapter 6) will tend

to obscure the distribution of edge lengths due to ancestral polymorphism (although this can be accommodated by an analysis of variance type approach). Also it is essential that loci have not undergone recombination, and are therefore a patchwork of informative sites with different histories. If so, then edge lengths of the reconstructed trees will tend to be distorted as mentioned in the previous paragraph and the variance of the coalescent time reduced (so underestimating ancestral polymorphism).

Another way is to just count how many genes (as one loci each) support each of the possible trees. This approach is fine if loci are not recombining and there are enough of them. If they are undergoing recombination then what is most likely is that they will all tend to support the species tree (if they are long enough). Consequently this method would be biased towards underestimating the true ancestral population size, especially if a condition of only counting genes which significantly supported one or the other trees is applied.

An implication of our estimate of ancestral population diversity is that it is consistent with our hominoid ancestral population being large for a substantial period from prior to the divergence of humans, chimps and gorillas until the divergence of humans and chimps (after which we have no direct knowledge). In this sense, the ancestral population would be more like that of modern chimps and gorillas, than our own species which shows signs of a recent expansion from a small (approximately 10,000) prior condition (Cann *et al.* 1987, Vigilant *et al.* 1991, Rogers and Harpending 1992). Evidence indicates that the subspecies of these species are the result of allopatric speciation and show no signs, as yet, of rapid expansions or sudden contractions in population sizes. In these large populations, non-neutral evolution tends to occur mostly through positive selection (so new behavioural or morphological features fixed at that time are expected to have been selectively advantageous). It would seem then, that the most reasonable hypothesis to explain the divergence of the human-chimp ancestral population from gorillas is that it was not unlike the separation of pygmy chimps from common chimps, for example. This speciation event appears to have been ongoing for well over a million years, and is marked by some relatively minor but noticeable shifts in morphology and behaviour. It seems reasonable to speculate that the first human ancestors may well have split from chimps in the same way (possibly 5 to 6 myr ago), but some time later began to experience accelerated evolution in what we consider distinctly human features. These features being bipedalism, increasing intelligence, more sophisticated tool use, fire, increasing socialisation (possibly marked by a reduction in male / female dimorphism), and eventually a shift into a more dominant ecological niche. This last event is marked by the appearance of ancestors as large as ourselves (e.g. *Homo erectus* and *H. ergaster*), which probably coincided with their ability to compete directly with the large carnivores, at about 2 to 2.5 myr ago.

This approach is certainly a method for the future. Over the next 10 years we are told to anticipate megabases of this type of data. Over regions of this length, the counts of sites supporting each of the alternative trees should be at least 10 times more accurate than what we have here, and there will be no doubt that recombination is indeed occurring at an appreciable rate up and down such as region (or even more so if it is a compilation of say 100 regions like the

β-globin locus). In other words, this method of analysis is awaiting the data which is coming. The other factor is that our estimates of t_{div}, t_{gen} are simultaneously improving. Accordingly, I expect that the solution to the question of how large our ancestral population size was at this time, will be answered definitively quite soon.

At present, extended research on this topic is being carried out with Dr Richard Hudson. My thanks to him for many helpful discussions on these topics. (Thanks also to the Smithsonian Institution for funding of stipend and travel during part of this research.) Our aim is to construct a reliable statistical genetic framework for all these related methods and approximations. Of particular interest is evaluating how different rates of recombination change the statistical properties (especially sampling variance) of estimates of P(incongruent) in going from completely linked to completely unlinked sites.

5.5 ROBUSTNESS OF TREE SELECTION IN THE FELSENSTEIN ZONE

The aim of this section is to assess the robustness of different tree selection methods to violation of the assumptions under which they are known to be consistent. We are particularly interested in the robustness attributable to the tree selection criterion, separate to the transformation of the data. The robustness referred to here, is the ability to recover the correct unweighted tree given sufficient sequence length; this is a form of consistency in spite of model violations. An alternative definition of robustness, is to evaluate how often a tree recovers the correct unweighted tree from a finite sequence length when the assumptions of a method are violated. And yet another definition of robustness is to define deviation from expected edge weights. Discontinuity introduced by the discreteness of trees makes such measures a little intricate, but deviation from expected path length is one general approach. The two latter definitions are used in one of the few papers to consider the properties of a tree selection method when the methods assumptions are violated (Fukami-Kobayashi and Tateno 1991). Concern for the recovery of accurate edge weights have their place (e.g. when trees are used to infer relative divergence times, Waddell and Penny 1995), but recovering the underlying tree usually takes precedence, even in such cases.

Robustness is a very important criteria because we all recognise that the process of evolution is more complex than our models. Indeed, when pushing phylogenetic inference to the extreme by using ancient molecules (older than 200 million years), the data are subject to strong selective constraints on sites and many misleading processes can occur. Earlier it was shown that sites in a sequence changing their relative rates would make the Hadamard method potentially misleading (chapter 2, appendix 2.1) due to inaccurate path length corrections. It is conjectured that the same model could mislead all the methods discussed earlier in this chapter (including likelihood), as well as more commonly used methods such as distance based tree selection. In chapter 3, we found a variety of methods (especially those based on transversional changes) which significantly rejected such well supported hypotheses as nematodes are Metazoa (= animals), while other methods based on the same sequences did not. Robustness then is a very desirable criteria, especially when working with deeply diverged functional molecules. Table 5.10 summarises the results of the next three sections.

The nomenclature for sections 5.5 to 5.6 is slightly nonstandard. The values returned by the i.r. Hadamard conjugation are here referred to generally as γ, and while in a real situation, tree selection is usually from sampled data, these are not strictly speaking sample data. Generally, unless it is to emphasise sampling properties, the symbol $\hat{\gamma}$ is not used in these sections.

Table 5.10. Summary table of the point of inconsistency (value of x) of different methods in the Felsenstein and anti-Felsenstein zones (when either too few or too many sites, respectively, are treated as invariant).

This table is meant as a 'lookup' summary only. All details of the models and the tree selection methods used are found in the following sections 5.5.1 to 5.5.3 and 5.6-5.6.5.

The method is a tree selection criteria followed by a form of data. Abbreviations are Pars. = parsimony, NJ = neighbor joining, all other method terms defined in text of appropriate sections. '+ve' means a tree is only considered if all edges in it are positive, '-ve' means -ve edges are allowed in the optimal tree. If a value of x is given, this is the point at which inconsistency will occur, if this is followed by unresolved (unres.) then inconsistency is avoided by the method defaulting from picking a resolved tree at this value of x. 'Same', means same as column to left, 'consist.' means consistent for all values of x examined, 'uncertain' means point of inconsistency is uncertain as it relies upon calculating the likelihood of site patterns given negative tree edge lengths.

	Felsenstein zone		anti-Felsenstein zone	
Method	+ve only	-ve allowed	+ve only	-ve allowed
Pars / obs.	$x < 0.045$	same	consist.	consist.
NJ / obs.	$x < 0.045$	same	consist.	consist.
NJ / δ	$x < 0.0375$	same	$x < 0.0215$	$x < 0.0215$
OLS / δ	$x < 0.0375$	same	$x < 0.0215$	consist.
Pars. / $\hat{\gamma}_D$	$x < 0.0375$	same	$x < 0.0215$	$x < 0.0215$
Pars. / $\hat{\gamma}$	$x < 0.0375$	same	$x < 0.0295$ unres.	$x < 0.0215$
GLS / δ	$x < 0.0350$	inconsist.	$x < 0.0215$ unres.	consist.
WLS / $\hat{\gamma}$	$x < 0.0302$	same	$x < 0.0295$ unres.	$x < 0.0135$
GLS / $\hat{\gamma}$	$x < 0.0310$	same	$x < 0.0215$ unres.	$0.016 < x < 0.0316$
ML	$x < 0.0235$	same	$x < 0.0245$ unres.	uncertain
min. X^2	$x < 0.0215$	same	$x < 0.0246$ unres	uncertain
$G^2\hat{\gamma}(T)$	$x < 0.0215$	same	$x < 0.0295$ unres.	uncertain
$X^2\hat{\gamma}(T)$	$x < 0.0155$	same	$x < 0.0295$ unres.	uncertain

5.5.1 Robustness to URAS of parsimony and neighbor joining

The model used here to assess robustness is the two state equifrequency mechanism of change, for which the Hadamard is proven to be consistent (Hendy and Penny 1993). In addition the sequences are modified so that they contain 50% of sites which cannot change (invariant sites). Under this type of model, most methods (with the known exception's of Lake's (1987) invariants and the Jukes-Cantor invariants of section 4.6.2), which are not modified to take into account unequal rates of substitution across sites, will suffer from a long edges attract form of inconsistency. This is effect is exacerbated by modeling the sequences on what has become known as a 'Felsenstein tree' (Felsenstein 1978a), (as shown in figures 2.1, and 5.11). The length of the long edges in the tree are fixed, but the length of the two other external edges plus

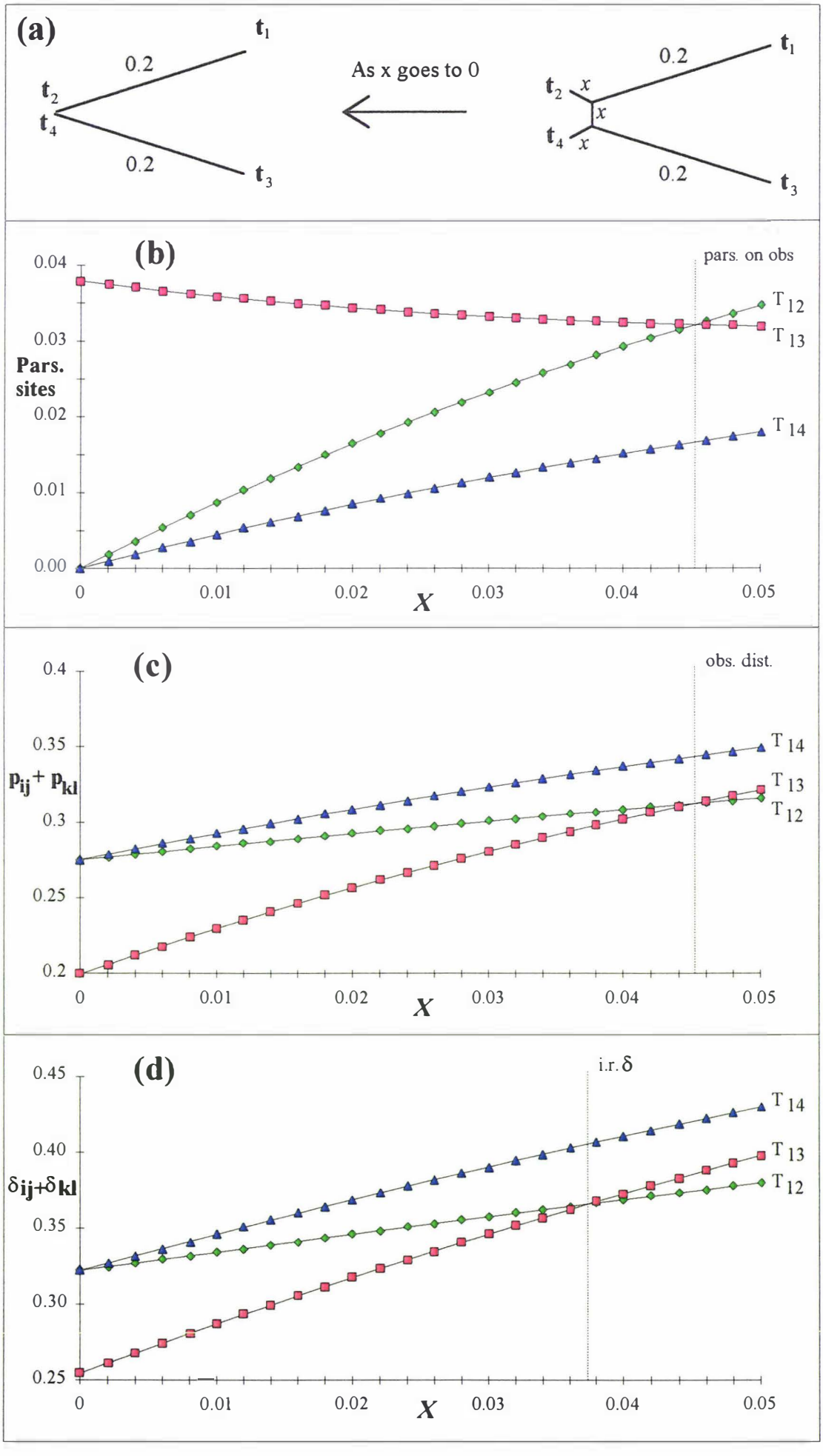

(a)
t1
0.2
t2
t4
0.2
t3
As x goes to 0
t1
0.2
t2
x
x
t4
x
0.2
t3
(b)
pars. on obs
0.04
0.03
Pars.
sites
0.02
0.01
0.00
T12
T13
T14
0
0.01
0.02
X
0.03
0.04
0.05
(c)
obs. dist.
0.4
0.35
pij + pkl
0.3
0.25
0.2
T14
T13
T12
0
0.01
0.02
X
0.03
0.04
0.05
(d)
i.r. δ
0.45
0.40
δij+δkl
0.35
0.30
0.25
T14
T13
T12
0
0.01
0.02
X
0.03
0.04
0.05

FIGURE **5.11a** The model tree used to generate sequences which will mislead i.r. / i.i.d. tree reconstruction methods as edges marked x tend to zero (edge weights are given as the expected number of substitutions per site, averaged across all sites). The mechanism for 2-state (e.g. transversional changes) character evolution is an equifrequency-Poisson process (e.g. Cavender 1978, Hendy and Penny 1989), except that 50% of sites are invariant, that is, unable to change. For the purpose of evaluating goodness-of-fit statistics we give the sequences a nominal length of 1000 sites. **5.11b** The proportion of sequence sites showing each of the three possible parsimony patterns (the total number will therefore be 1000 times this). The dotted vertical line indicates when parsimony applied to these data will become inconsistent. **5.11c** Lengths of pairs of observed or uncorrected distances, $p_{ij} + p_{kl}$ (where i, j, k, l index each of the four sequences) inferred for each tree. Many distance based tree building methods, including neighbor joining, minimum evolution, and the ST method will always chose this tree under this model, when applied to uncorrected distances. **5.11d** The expected value of $\delta_{ij} + \delta_{kl}$, where δ is the "corrected" distances under the i.r. / i.i.d. 2-state Poison model. These methods are more robust than those in 5.11b or 5.11c, but also become inconsistent for $x < \approx 0.0375$

the internal edge are made to take value "x". Fixing the long external edge's length, limits the length of the longest path through the tree at the "inconsistency point". For this model, the inconsistency point requires $x < 0.05$, so the longest path is always less than 0.5 changes (averaged across all sites, including the sites which cannot change). We have chosen this model (rather than one where the long edges are allowed to increase in length) to counter any argument that inconsistency is only prominent when using sequences which have diverged by such an amount that biologists should not be using them for phylogenetic inference. Tree selection criteria are evaluated on the three possible binary trees, and when illustrative, also on the star tree.

Figure 5.11b-d shows the consistency of commonly used methods which do not optimise a standard statistical criteria of fit of tree to data. Figure 5.11b shows the frequency of each of the parsimony patterns, all of which show near linear trends as $x \rightarrow 0$. In this example, parsimony will always pick the tree supported by the largest number of informative sites (i.e. the highest y value). That is, parsimony, or compatibility, applied to this set of observed 4-taxon sequences will always converge to the tree for which the pattern in s(T) corresponding to an internal edge is most frequent. For all values of x left of the dotted line ($< \approx 0.045$) these methods are inconsistent, i.e. they will always choose an incorrect tree given sufficiently long sequences. This is a somewhat larger value of x than would be the case for the same model without invariant sites (where x must be $< \approx 0.028$), for example, the model used by Felsenstein (1978a) to first show the inconsistency of parsimony and compatibility. Because we are measuring the average rate across sites as including the 50% of invariant sites, then the expected length of each edge measured on just the variable sites is twice that shown in figure 5.11a. Consequently, the inconsistency point of parsimony in this model is the same as that for parsimony under the i.r. 2-state Poisson model when the long external edges are 0.4, and x is $\approx 2 \times 0.045$ or 0.09. A surprising feature as x tends to zero, is that while patterns supporting the true tree T_{12} (and also T_{14}) go to zero, the proportion of misleading parsimony patterns continues to rise.

Figure 5.11c shows the expected value of the three pairs of observed distances (p_{ij}) such that each taxon is only used once in a pair. Each pair, by Buneman's (1971) 4 point metric condition, is associated with a particular binary tree. Algorithmic tree building methods such as neighbor joining or the ST method (Sattath and Tversky 1977, Charleston *et al.* 1993) applied to four taxa, will always choose the tree for which $p_{ij} + p_{kl}$ is minimal. Accordingly, for figures 5.11c-d the optimal tree at a particular value of x, is that with the smallest y-axis value. The point at which this combination of data correction (none in this case) plus tree selection procedure becomes inconsistent on this data is exactly the same point at which parsimony becomes inconsistent. A quick algebraic exercise will show that the only difference between the value of $p_{ij} + p_{kl}$ on different trees, T_{mn} and T_{op} is $-2 \times$ (the number of parsimony patterns supporting T_{mn} - the number of parsimony patterns supporting T_{op}), since the only difference in observed distances, are the patterns corresponding to informative parsimony sites. This identity between the point of inconsistency of parsimony and neighbor joining applied to 2-state, or 4-state data is in agreement with the proof in Penny *et al.* (1991). The y-axis values of figure 5.11c gauge approximately what the overall sum of branch length in the reconstructed tree will be; at its worst (fig. 5.11c, T_{13}, $x = 0$) the reconstructed sum of branch lengths is only half the total number of changes that actually occurred.

Figure 5.11d shows the value of $\delta_{ij} + \delta_{kl}$, associated with each binary tree (where δ is the i.r. model corrected distance). As noted above methods such as neighbor joining and ST applied to this model data will chose the tree for which $\delta_{ij} + \delta_{kl}$ is minimised. Notice, that the value of x for inconsistency to occur has become smaller ($x < \approx 0.0375$) (and thus more extreme) with this prior transformation of the data. Accordingly, we say that this combination of transformation followed by tree selection combination is more robust than the previous combination of with no transformation (all of these former methods being equivalent to parsimony on the observed data in this instance). An obvious conjecture is that this result will extend to larger trees where sites evolve by the same mechanism (50% invariant, remainder 2-state i.r.). This of course is not trivial to prove. Note also, that this result is not a general conclusion for all models, since in section 3.4.3 we have already shown a logarithmic transformation leads to increased inconsistency under some non-stationary base frequency models.

With four taxa and 2-state sequences, any of the parsimony-compatibility family of tree selection methods (Felsenstein 1981b) applied to either identical rates $\hat{\gamma}$ or i.r. $\hat{\gamma}(D)$ will the tree minimising $\delta_{ij} + \delta_{kl}$ (see for example Steel *et al.* 1993b). Consequently, the inconsistency point for neighbor joining, minimum evolution, ST and the like, is here shared also by unweighted, i.i.d. tree selection applied to the i.r. spectra of Hendy and Penny (1993). An obvious conjecture to make at this point, is that all these criteria are equally robust for this type of mechanism of substitution. As mentioned later in this chapter, that this is not the case. Notice that in the worst case in figure 5.11d (T_{13}, $x = 0$) we are still dramatically underestimating the sum of edge lengths in the true tree by $\approx$ 1-0.25/0.4 or $\approx$ 37%.

5.5.2 Robustness of WLS methods of tree selection from $\hat{\gamma}$ and δ

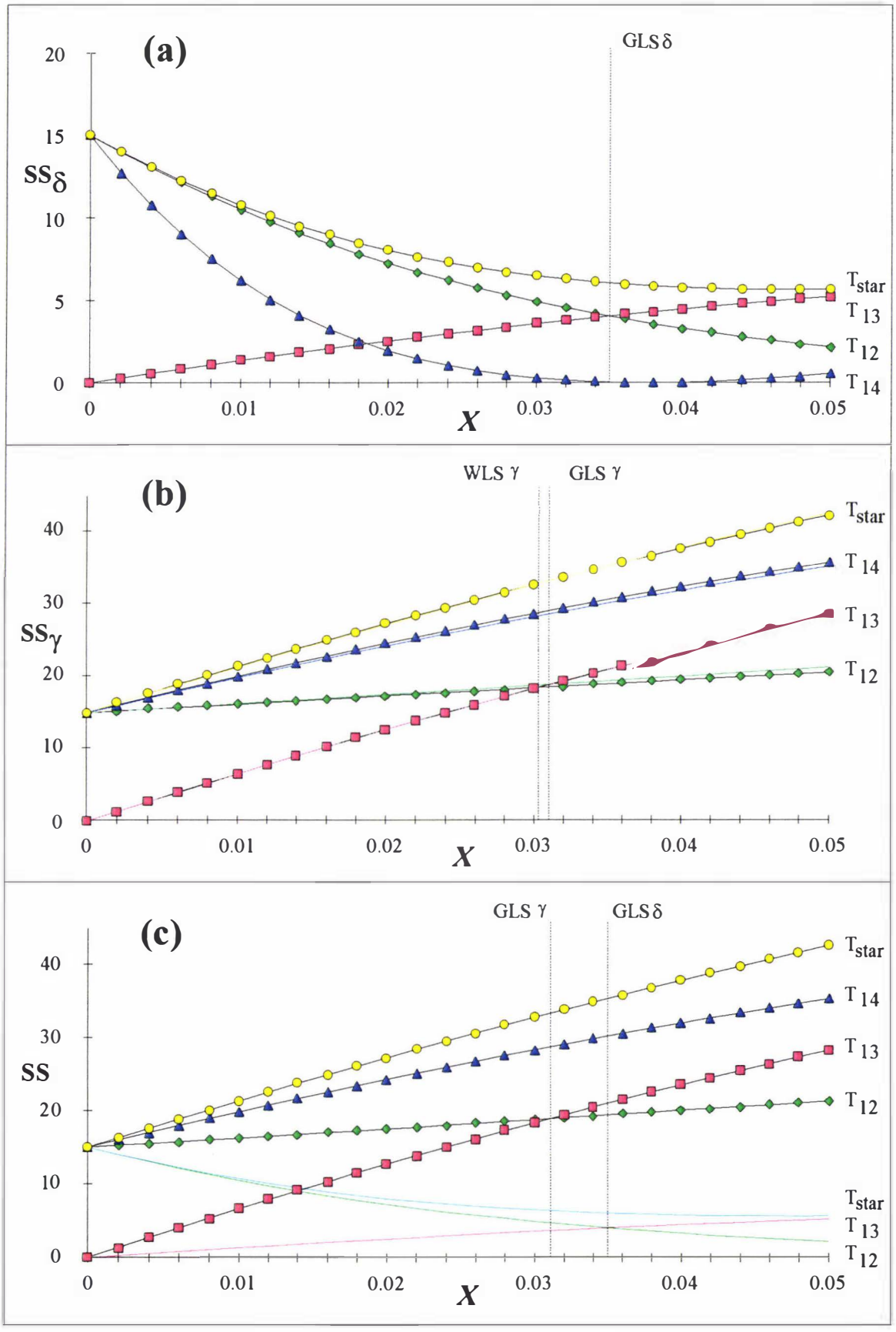


FIGURE **5.12a** A plot of the GLS sum of squares (SS) fit for the model data to each tree, when working from just the 6 pairwise distances. Note that in this situation, T_{14} has a negative internal edge length.

5.12b The WLS sum of squares (SS) fit for the model data to each tree, when working from the entries in γ corresponding to possible edges. Next to these values are the nearly identical GLS SS from γ shown as lines without symbols. In the case of T_{13} and T_{star} the differences are all but indistinguishable. For T_{12} the GLS values are slightly higher than the WLS values, while the reverse is true for T_{14}. **5.12c** The GLS SS fit of the model data to each possible tree when measured on both γ (lines with symbols) and distances, δ (lines without symbols), shown to scale.

In this section, the robustness of more "sophisticated" and computationally expensive statistical methods of selecting the optimal tree are evaluated. Figures 5.12a-c show WLS and GLS methods applied to γ and GLS applied to δ (the i.r. transformed distances). These graphs show the fit between the transformed sequence data and each tree optimised, so as to maximise these fit criteria (using the equations in section 5.2.5 and 5.2.6). The fit (sum of squares, SS) is scaled to represent a sequence of 1000 sites. This value does not take into account the fluctuations in the SS due to sampling error (i.e. finite sequence lengths). This error is expected to be non-centrally chi-square distributed (Stuart and Ord 1990, p. 865) for long sequences, and will almost always inflate the SS that calculated here (with the expected amount of this inflation tending to be near equal for the different trees).

Figure 5.12a shows the results of GLS tree selection applied to distances for the three binary trees and also the star tree. This method becomes inconsistent for x less than 0.035, which is a slight improvement over the robustness of the previous classes of methods. A prominent feature of GLS applied to this data is that the internal edge of T_{14} has already collapsed for $x > 0.05$ and has a negative edge weight, making its SS appear much better than it should be. One logical approach outlined at the beginning of this chapter, dictates that negative edge lengths be set to zero, and a re-optimisation done (the logic for doing this is extensively discussed in Felsenstein 1984, while the practical argument of increased convergence is supported by results in Kuhner and Felsenstein 1994). With this constraint in place the SS of T_{14} on this data becomes identical to that of T_{star}. An alternative approach, also suggested at the beginning of this chapter, is that any resolution involving a negative edge be ignored, if there is a tree with a positve internal edge length, else chose the tree with the least negative internal edge. With either approach, this method does not become inconsistent until $x < 0.035$. Otherwise, choosing the minimum GLS SS irrespective of internal edge length in long edges attract situations, will probably result in a reduction in the accuracy of tree estimation.

These results also show that GLS is more robust than neighbor joining or minimum evolution under our model. This finding tempers the claim of Rzhetsky and Nei (1992b) that the minimum evolution method is superior to GLS (on i.r. corrected data); just because it has claimed faster convergence under a simple model with all assumptions meet does not make it generally superior.

On a different tack however, figure 5.12a shows evidence of limitations of GLS tree selection on distances, relative to other statistical methods, as suggested by the analyses of the four taxon rRNA data (performed earlier in section 5.2.8). Specifically the GLS applied to δ SS in figure 5.12a gives little indication that the data is really rejecting the model for any value of x (as would

be measured by the chi-square distribution of SS expected when the null model is true). Worse still, the main region where the model is expected to fits acceptably is in the zone of inconsistency, and on the incorrect 'long edges attract tree' (as x drops below approximately 0.01)! This result serves to show, yet again, that overall fit of data to model cannot be relied

upon to tell us if tree selection is reliable (we discuss this in more detail in chapter 6). The GLS SS measure of fit between transformed data and optimal tree, will increase linearly with the sequence length specified (irrespective of the form of the data, i.e. δ or γ, and at least as c becomes large). That is, if we had a sequence of length twice that specified here (i.e. 2000), then the SS for all trees in figure 5.12a is expected to be double that shown at all values of x. As we have 1 degree of freedom (d.f.) for our residual, our critical chi-square value for rejecting the null model (with $\alpha = 0.05$) is 3.841, so for 2000 base pair sequences we would be rejecting the null model most of the time as long as $x > 0.01$. Its difficult to say exactly how often rejection will occur without specifying the non-central chi-square distribution of the SS for each tree-model. That is, this distribution is not fixed for each tree, but rather the non-centrality parameter will change as x changes. What is perhaps more important, is the comparative performance of this method which is discussed further below. The relative insensitivity of GLS on distances must be largely due to having lost information in going from sequences to distances, since GLS on γ is much more sensitive, as is shown next. These results are consistent with those in section 5.2.9, where GLS on distances did not have the discriminatory power of GLS on transformed sequences.

The next figure 5.12b plots the WLS (lines with symbols) and GLS (associated lines without symbols) for different trees evaluated on i.r γ. Both methods have made a considerable gain in robustness over the previous estimators, which included closest tree, parsimony etc. on γ. A striking result is just how close the two sets of SS are, which implies that overall the covariances in γ are having a small effect on the SS, even though this type of tree-model has some notable correlations (see chapter 4 and earlier sections of this chapter). Part of the reason will be due to the covariance matrix estimated under the i.r. model underestimating the rates of change, and so too the correlations (since the pathlength transformation is less severe than it would be if invariant sites are taken into account). Interestingly, and promisingly (since we don't relish the computational complexity of GLS vs WLS for tree selection) WLS is here more robust than GLS. For a sequence of length 1000, both methods applied to γ will detect the mismatch of data and model as long as $x > 0.006$. Generally GLS on γ shows a vast improvement in power to reject as inadequate the fit of data to model, relative to GLS on δ. The extra degree of freedom for the GLSγ SS makes little difference to this impression, since it only raises the required rejection level at $\alpha = 0.05$ from 3.841 to 5.991. This striking difference in the ability of the SS on γ vs δ to detect deviations from the model is shown in figure 5.12c, where both these GLS SS are plotted to the same scale. Interestingly in the limit as $x \rightarrow 0$, both methods not only return equivalent sums of squares but moreover become pretty much identical. This is understandable since, when

$x = 0$, all trees are effectively 3-taxon trees (with t_2 and t_4 equivalent) so all the paths in the Hadamard conjugation are reduced to just pairwise distances.

5.5.3 Maximum likelihood is inconsistent when there are invariant sites

Next, we consider the performance of four tree selection criteria which measure the goodness-of-fit between observed and expected data using well known statistics at the **s** level. The optimised methods are maximum likelihood (ML) via the minimum G^2 statistic and minimum X^2 (sometimes called minimum chi-squared). In addition there are two unoptimised fit methods as described in section 5.2.10. The first is the likelihood (or X^2 statistic) associated with each tree when edge weights are taken directly from $\hat{\gamma}$ (called $L\hat{\gamma}(T) = G^2\hat{\gamma}(T)$) or the unimproved X^2 fit of each tree $X^2\hat{\gamma}(T)$ (depending on the statistic used).

Looking firstly at ML, figure 5.13a shows some very promising results. ML has substantially increased robustness over the best of the previously considered methods (GLS on γ), becoming inconsistent only when $x < 0.0235$. In addition its fit statistic G^2, is about 25% bigger than that for GLS on γ, making rejection of the null model even more likely, since under the null model both distributions are expected to be chi-squared with 2 d.f. Otherwise the statistics GLS SS on γ and the likelihood ratio statistic change in a very similar manner. With ML, rejection of the null model (that the data came from this tree model) only becomes unlikely for $x < 0.004$. Notice how close the fit of the star tree is to T_{14} in figure 15a, a change from GLSγ were T_{14} had a noticeably lower SS then T_{star}, (a possible trend noted earlier in section 5.2.6 with the 2-state rRNA data whereby GLS on γ gave more support for an internal edge than did ML or iterated GLS).

This demonstration of the inconsistency of ML when based on a mechanism of change (or model), which will increasingly underestimate the actual number of substitutions occurring as the true amount of evolution increases is novel. As explained later (section 5.10), others have suggested inconsistency will occur (based on performance in simulations with finite sequence lengths), but no one has shown it with precise calculations. This finding also serves to generalise the phrase of Hendy and Penny (1989) 'long edges attract'. We see no reason why this phrase should be restricted to describing the behaviour of parsimony on observed sequences (as used originally by Hendy and Penny 1989), which is just one of many methods which underestimate the true number of changes with a homogeneous Markovian mechanism of substitution when rates at different sites are unequal. To complete this generalisation of 'long edges attract' we must show that a tree which can be rooted and have all external tips equal (i.e. a tree consistent with a molecular clock) will also cause likelihood to be inconsistent (not just an unequal rates effect, as considered by Felsenstein 1978). We evaluate the performance of ML and X^2 in the 'Hendy-Penny zone' in a later section of this chapter.

The performance of the uniterated likelihood method, $L\hat{\gamma}(T)$ (here represented as $G^2\hat{\gamma}(T)$) is shown in figure 5.13b as the line with symbols (also shown as plain lines are the iterated maximum likelihood = minimum G^2 values). Yet again the results are surprising and pleasing,

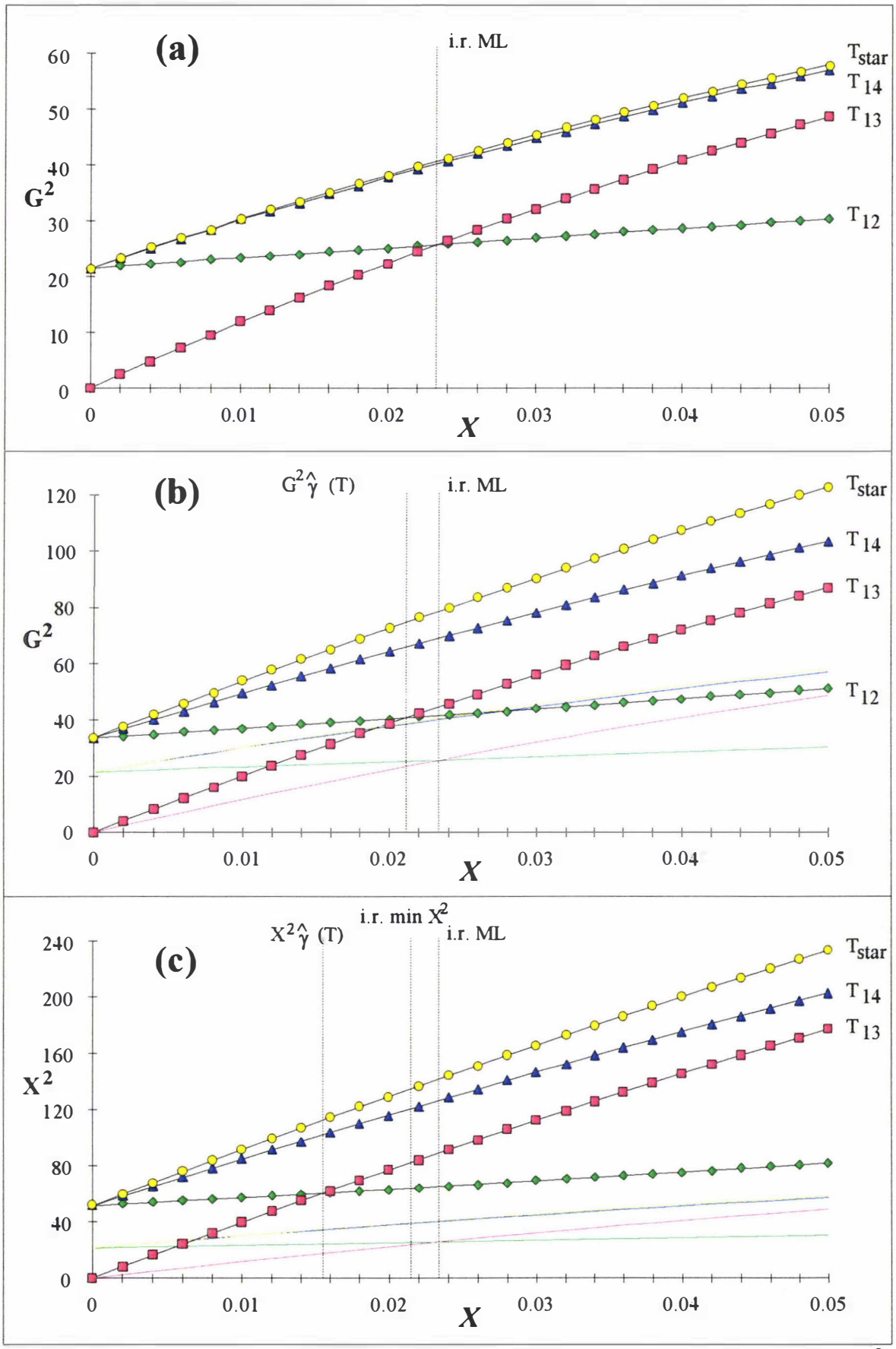


FIGURE **5.13a** The fit of between each ML tree and the model data as shown by the likelihood ratio statistic G^2. **5.13b** The lines with symbols show the uniterated G^2 statistic for the fit between the observed data and the data predicted by each tree with edge weights taken directly from $\hat{\gamma}$ (i.e. $G^2\hat{\gamma}(T)$). Lines without (continued)

symbols show the G^2 statistic for the same trees with edge weights chosen so as to minimise G^2 (i.e. the ML estimates). **5.13c** Similar results for X^2. The lines with symbols show the values of $X^2\hat{\gamma}(T)$, the lines without symbols are the values for ML from 5.13a, and the just point of inconsistency for minimum iterated X^2 is shown.

this time the method $G^2\hat{\gamma}(T)$ achieved an even better degree of robustness than ML. We don't yet know if this result will generalise to other trees, but it may be an example of a hybrid tree selection criteria having especially desirable features. A possible analogue of a hybrid phylogenetic method having some clear advantage over an a well established statistical method is the minimum sum of OLS edges -minimum evolution method (Kidd and Sgaramella-Zonta 1971, Rzhetsky and Nei 1992a). Like minimum sum OLS edges, our $G^2\hat{\gamma}(T)$ method constructs edge lengths by an OLS analogue (the uniformly weighted Hadamard transform), and then evaluates the exact likelihood of each alternative tree for these edge weights (rather than using sum of absolute deviations). Certainly both the robustness of $G^2\hat{\gamma}(T)$ (as well as its convergence properties) need further investigation and cannot be taken for granted from 4-taxon examples.

Figure 5.13b also shows that while the trends in $G^2\hat{\gamma}(T)$ mimic those of the ML estimates, they can give goodness-of-fit statistics which are twice as large or more. As yet we have no formulae to predict what the expected distribution of $G^2\hat{\gamma}(T)$ will be under an incorrect model. Asymptotically, under the true model, its fit could be far worse than a chi-square variable with $n-1$ degrees of freedom (where n is the number of different sequence patterns). This is because it is known that less efficient estimators than ML (including methods of moments) can lead to χ^2 sampling distributions with more degrees of freedom than n-1 (Stuart and Ord 1990, p. 1171), and it seems likely that with more taxa, edge length estimates taken from $\hat{\gamma}$ are statistically inefficient compared to conditioning on a single tree. Thus we know that the expected value of $G^2\hat{\gamma}(T)$ is larger than that expected of an ML estimate (under the null model). Since the expected $G^2\hat{\gamma}(T)$ is so different to the iterated ML minimum G^2, then we must be extremely cautious about using a statistic $G^2\hat{\gamma}(T)$, in combination with a χ^2 distribution, for statistical testing (more on this in chapter 6).

The last criterion considered is use of the Pearson X^2 statistic as a measure of fit of tree-model to data. We will use X^2 both with explicit minimisation (hence called minimum X^2), and as a measure applied to the s(T) vector of trees whose edge weights come straight from $\hat{\gamma}$ (this being labeled $X^2\hat{\gamma}(T)$, an uniterated form of minimum X^2). The results in figure 5.13c indicate that tree selection by the X^2 statistic is as robust (and perhaps slightly more so) than tree selection by the G^2 statistic. The iterated minimum X^2 tree selection method was more robust than G^2, requiring x to be less than 0.0215 before inconsistency occurred (see figure 5.13c). The trends for the minimum X^2 criterion are similar in shape to those for minimum G^2, but as with the 4-taxon 2-state rRNA example, X^2 appears a more powerful statistic for detecting this type of deviation from the null model, taking values approximately 12% larger than G^2 for T_{12} near x = 0.02, and even larger for the other trees (approximately 20% bigger than T_{13}). If this increased power is a general feature of X^2 in detecting the type of departures generated by inadequate

correction for multiple changes, then it may well be worth implementing minimum X^2 alongside ML methods. In the case of ML via Hadamard conjugations minimisation should only require a few extra iterations, as the ML minima is usually close to the X^2 minima (its implementation by traditional methods is slowed by need to evaluate many more probabilities). It need not be applied to all trees, perhaps only helping to distinguish amongst the subset of the best trees selected by ML.

The results for X^2 applied to $\mathbf{s}(T)$ for trees with edge weights from $\hat{\gamma}$ are even more encouraging than those for minimum X^2, with x needing to drop below 0.0155 before inconsistency occurred (figure 5.13c). However, as with $G^2\hat{\gamma}(T)$, it is clear that the residual X^2 is markedly larger than that of minimum X^2 and we must be cautious about using this statistic to measure fit of data to model without knowing more about its sampling distribution. One advantage of X^2 over G^2 as a test statistic, is that the deviation per cell is obvious, with terms all being of the same sign and, additionally, each cell has an approximately chi-squared distribution with 1 d.f. (under the null model, with long sequences and d.f. approaching the number of cells). By examining X^2 on $\mathbf{s}(T)$ (not shown) it was apparent that for all trees the major contribution to the large X^2 statistic (well over 95% of the total) were the two cells corresponding to the two non-tree parsimony patterns. All the other cells had surprisingly good fits, with only that of s_0 often being greater than 1, and then typically contributing no more than 3 % to the total X^2 statistic (a somewhat counter intuitive result since the actual model is violated by having many more constant columns than expected). These findings applied for all the values of x considered. Thus, the X^2 statistic seems to be particularly sensitive to rarer patterns which have a higher than expected frequency, which is anticipated when more substitutions are occurring in reality than our tree plus mechanism of evolution predicts. This helps to explain the increased robustness of these criteria to 'long branches attract' problems. Just the converse may be the case for the data in section 5.3.7, where overcorrection might be a problem, and X^2 is possibly less sensitive to deviations from expected fit than G^2 (see also section 5.3.8).

5.5.4 Inconsistency with a continuous distribution of unequal rates across sites

Some readers may feel that the previous example was somehow artificial because we had an extreme of variable and completely invariable sites. This is not the case, and here we show that the same type of inconsistency (as shown above with 50% invariant sites) occurs for any fixed non-i.r. distribution of rates across sites if the model assumes i.r. In the case where rates across sites follow a Γ distribution with shape parameter $k = 1$, the effects are very similar to what we see when there are 30% invariant sites (the pathset transformation curves in figure 2.6 makes a similar prediction). For all of the tree selection procedures, the edge length x had to be smaller than in the 50% invariant sites model to cause inconsistency, but inconsistency still occurs. Figure 5.14 shows how parsimony on observed sequences, parsimony on i.r. γ, and i.r. ML fare when sequences are generated with a Γ distribution of rates across sites ($k = 1$). The overall trends are very similar to the case with 50% invariant sites, but for the smaller values of x at which inconsistency occurs (especially for i.r. ML for which x must be over twice as small before

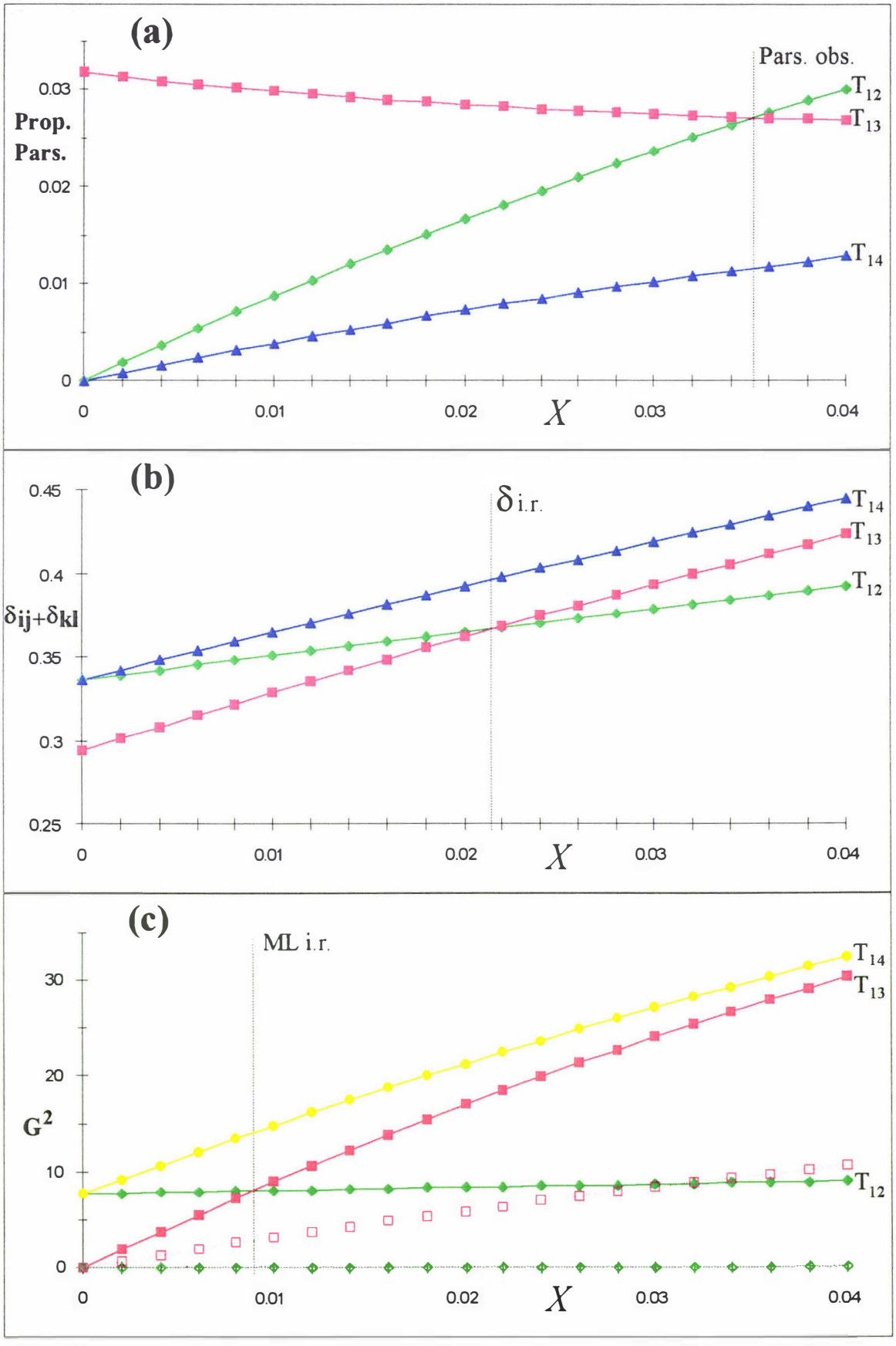


FIGURE **5.14** The robustness of three methods when an identical rates model is assumed, but site rates follow a Γ distribution with shape parameter k = 1(the mechanism of evolution is 2-state Poisson, while the tree is as shown in figure 5.11a) and there are no invariant sites. (continued)

5.14a Parsimony on observed patterns. **5.14b** Parsimony on i.r. γ (here equals neighbor joining on i.r. model corrected distances). **5.14c** Maximum likelihood under the i.r. model (solid symbols indicate fit of trees under the i.r. model, while the hollow symbols with dotted lines indicate their fit under the i.r. + optimised p_{inv} ML model).

inconsistency occurs). ML in combination with the G^2 statistic shows a good ability to reject the data coming from the assumed model, except when $x < \approx 0.003$.

Figure 5.14c also shows how the i.r. + invariant sites ML model performs on data generated under a Γ distribution ($k = 1$). The hollow diamond shaped symbols are the fit of the true tree T_{12} with the i.r. + p_{inv} (invariant sites model), while the square symbols are the fit of T_{13} (actually the same as the fit of the star tree most of the time, and otherwise indistinguishable at this resolution). Clearly the fit of T_{12} is near perfect, while all other trees meet the expectation that under a well fitting model, they should have no more support than the star tree if they do not share internal edges with the tree generating the data. We evaluated the relative G^2 statistics of the different trees at a value of x as small as 0.0001 (about 0.3mm away from $x = 0$ on the scale of the figure, or 1/6 the width of a square symbol). The wrong tree T_{13} was still worse fitting than T_{12}, so if i.r. + p_{inv} ML is inconsistent in this situation, it is at a value so small as to be practically meaningless.

Thus we conclude, that here at least the i.r + p_{inv} model is consistent. This finding reinforces our expectation that the i.r. + p_{inv} ML method (and similarly distance methods also) will have good robustness against problems due to a stationary unequal distribution of rates across sites (with respect to unweighted tree selection but not necessarily edge length estimation). Further the difference in goodness-of-fit between the invariant sites model and the Γ distribution ML method was so small that the difference could not be seen at the resolution of figure 5.14c (i.e. their lines exactly overlay one another).

One last feature of figure 5.14c is also striking. Notice how the fit of T_{13} (= fit of T_{14} = fit of T_{star} in this instance) has become much better itself. Accordingly the difference in likelihood (or fit) between the best tree (here the true tree) and incorrect trees has dropped substantially over much of the range of x. This echoes a finding made in the study of hominoid mtDNA for Penny and Waddell (1995, see section 5.3.9), which has is also noted by Yang *et al.* (1994), where the difference in fit between trees can drop significantly when allowing for a variation of rates across sites (note however that in figure 5.6a-c this does not always happen).

This model-based study shows that the convergence in likelihood of different trees when taking account of URAS need not be artifactual, due, for example, to the simplifying assumptions made in order to place likelihoods upon the real process of evolution. It can be understood theoretically in the asymptotic context of iterated GLS converging to ML. As rates across sites become larger, section 4.5 shows that variances and covariances will rise, thus the GLS SS difference between trees will often decrease. Of course, the opposite can also happen; consider a situation where two trees are nearly tied for optimal, but when rates across sites are taken into account the support for one decreases.

However, across all trees, and in most situations, the expectation is that the increasing variances and covariances will decrease resolution. Another way of looking at the same thing, is that with the larger amounts of change implied by taking account of rates across sites, the predicted s(T) of different trees become more similar as they both tend to predict a fair number of convergences and parallelisms. These give many cells higher expected values, especially important amongst the critical patterns which have a much higher likelihood on one tree verses another (often because a change on the internal edge of one tree, otherwise requires more than one change on the other tree). Because the X^2 and G^2 statistics tend to discriminate most strongly when something is observed in numbers when it should have been very rare, the discrimination between trees drops. (An argument can also be made for differences in raw likelihoods; all of these arguments are in essence correct, and only vary in their illustrative power). This of course is reality, and again raises the question of whether data editing will sharpen the resolution and improve the separation between trees (as section 4.5 shows, this must happen in some cases).

In conclusion to this section, we have seen an interesting trend whereby the selection criteria which can discriminate that the data does not come from the model, using well proven multinomial or multivariate normal fit statistics (which are also generally more computationally expensive) do have greater robustness than other tree selection criteria. It is these well known statistical criteria, which without modification, give a statistic of goodness-of-fit of data to model. These two attributes make a criterion such as maximum likelihood look favourable for dealing with processes of sequence evolution where we expect the greatest violations of our model, including a distribution of rates across sites, base compositional changes, concerted changes in groups of sites, sites undergoing substitution by more complex Markovian models than assumed, and different sites obeying different Markovian models. These are just the problems of homologous sites, not to mention artifacts of alignment (e.g. Lockhart *et al.* 1995, Olsen and Woese 1989). All of these problems are typical of rRNA and, the deeper the divergence, the more they are expected to come into play. However, even tree selection criteria with enhanced robustness can be swamped by deviations from the model (e.g. where ML based on homogeneous models, consistently puts *Giardia* branching earliest amongst the eukaryotes, other possible examples are furnished in Lockhart *et al.* 1995). Accordingly we should not give the criteria, separate to the assumptions of the mechanism of change, magical properties of always being correct.

Overall, the methods of minimum G^2 (ML), and the computationally more expensive minimum X^2, clearly did very well in this study in comparison to other criteria. The methods of uniterated ML and X^2 did well also, but require more study before we can feel confident of their general robustness (an area for further study being which method of estimating edge lengths gives the best performance). Also worth bearing in mind is the suggestion from chapter 3 that there may be situations where distance based methods, based on the same mechanism as ML, may be more robust. Our expectation is that this will not be a general property, but none the less it would be desirable to know in which situations each method offered the best robustness.

5.5.5 Different trees can give identical sequences!!

It has been one of the underlying assumptions of most phylogenetic methods that each tree generates a different set of sequence probabilities, so each tree can be uniquely identified in expectation (e.g. Barry and Hartigan 1987a). Here, we show that this need not be the case when rates across sites are unequal. The implications of this for tree estimation are then discussed. The realisation that different trees might give the same sequences when rates across sites varied was inspired by the way in which the pathset length correction curves with invariant sites and a Γ distribution can cross each other twice (excluding being equal at zero, see figure 2.6).

There is an example of two trees giving the same sequences in all the last series of figures 5.11-5.14. It is the terminal point in the Felsenstein zone. Notice how an incorrect tree, T_{13} not only becomes optimal by ML, but also perfectly fitting. Given that we have calculated the probability of all patterns exactly, this can only occur if a sequence for T_{13} can be exactly the same as a sequence for T_{12}. In this case the trees are slightly unusual, T_{12} under the a non i.r. 2-state Poisson model is ((1: 0.2, 2: 0.0): 0.0, (3: 0.2, 4: 0.0)) (a star tree with a long edge between 1 and 3, and the taxa 2 and 4 indistinguishable from a single node midway on this tree, written in a nested form, e.g. see Felsenstein 1993). Taking the sequences of this tree and analysing them assuming an i.r. distribution of rates across sites, gives a tree binary T_{13} ((1: y, 3: y): x, (2: 0.0, 4: 0.0)) (which has two adjacent external edges equal to zero, and where the x and y are determined by the distribution of rates across sites in the tree model for T_{12}). The length of y will always be less than 0.2 (the more severe the distribution of rates across sites in the first model the more so), and x will always be greater than zero (and getting bigger as the distribution of rates across sites becomes more severe, but never as big as y).

The reason these two trees have the same sequences as long as rates across sites can vary is understandable in terms of the Hadamard conjugation. Going from $\gamma(T)$ to $\mathbf{s}(T)$, the different models use different transformation curves going from ρ to $\mathbf{r}$, and with certain weighted trees and certain distributions of rates across sites it is possible to map the slightly different pathset lengths of ρ onto an identical $\mathbf{r}$ vector (hence onto the same $\mathbf{s}$ vector). We use the term 'slightly different' because since both transformations are monotonic functions, all pathsets for two different trees must fall in the same rank order in order to get this 'dual mapping' effect. Further detail on this interesting finding is given in appendix 5.1 where we present and discuss our results, and the independent result of Steel *et al.* (1994) which is an existence proof (which says something will happen, but does not describe where or how). No concrete example of two resolved binary trees mapping to the same sequences is yet known; it should not be too difficult, but does require finding two distributions of rates across sites where the correction curves (like figure 2.7) will cross at least six times (work on this is in progress). Results in appendix 5.1 suggest that as the number of taxa grows, it becomes increasingly unlikely to get exactly the same sequences for two trees, and to 'dual map' trees must have increasingly small internal edges. However, it does show a lot of 'Felsenstein' and 'anti-Felsenstein' effects in an extreme form; so extreme that without some knowledge of the distribution of rates across sites independent of the

sequences in the tree, one could not have any confidence in which was the true tree (even with infinite sequences, and perfectly fitting models).

These findings also require some further precision in the use of phylogenetic terms. In order to be able to say that additivity guarantees the consistency of tree selection from distances (with an appropriate algorithm), we must define additivity as being on the true, generating, tree (that is, true tree additive). Alternatively, the term additivity could be defined in a specific sense, meaning that for any collection of taxa (hence any weighted tree) additivity of distances will always hold. We use additivity in this sense within this thesis, and if we mean 'local' or 'observed' additivity, this is specified. While the use of the term additivity is important in proving the consistency of tree estimation from distances, there is a parallel for sequence based methods, including likelihood. That is, perfect fit of data to model only guarantees consistency if it can be shown that no two trees can give the same sequences. In appendix 5.1 some examples are given of models where we do not expect different trees to give the same sequences.

The situation as x goes to zero in our model (figure 5.11a) also yields a foundation for a mathematical proof that likelihood will be inconsistent. As x goes to zero, the vastly more frequent parsimony pattern will be that of long edges attract, and not that of the true tree. With all sites analysed being i.r. and i.i.d. then ML based on the i.r. / i.i.d. model will remain consistent. However, when invariant sites are added this upsets the balance. If we then add an increasingly large proportion of invariant sites, then the data will take on the characteristics of an infinite sites model, under which ML chooses the most frequent of the three possible parsimony patterns (e.g. Goldman 1990), and becomes inconsistent.

5.6 ROBUSTNESS OF TREE SELECTION CRITERIA IN THE ANTI-FELSENSTEIN ZONE

Here we introduce the concept of a new region of inconsistency of tree building which we call the 'anti-Felsenstein zone' (that is, the opposite of the Felsenstein zone, introduced in Felsenstein 1978). The anti-Felsenstein zone, is a case of 'long edges repel', which occurs when there is over-correction for multiple hits, combined with a certain type of tree generating the data (we call this an *anti-Felsenstein* tree). While it is a good idea to take into account the extra unseen multiple hits expected when rates across sites vary, it is also unwise to make too much correction. Jin and Nei (1990), for example, recommend using a distance correction for functional regions which assumes that the distribution of rates across sites is exponential, and not something less dispersed (while the program MEGA of Kumar *et al.* 1993, uses a default of k = 0.5 with some of its Γ distributed distances). Lockhart *et al.* (1994) went further, and excluded all constant sites (and sometimes even all singletons as well) in an attempt to overcome the effect of some sites being invariant (or at least very unlikely to change).

To date, no study has been made of what biases "overcorrection" of the data might introduce to tree estimation procedures. Olsen (1987), notes there may be a problem based on his studies with ancient rRNA, but does not expand on this supposition. Here we study the effects of making corrections under models which expect more multiple hits than actually occurred in the data, making quantitative predictions of what can go wrong, and which tree selection criteria appear to be the most robust to this problem.

An interesting feature of this zone, is that it often involves the inference of negative internal edge lengths. Whether or not these are considered valid, makes a large difference to the consistency of methods, with their exclusion tending to be the more appropriate option. This adds another dimension to the exchange of Farris (1981, 1985, 1986) and Felsenstein (1982, 1984, 1986, 1988) on this matter, and tends to support Felsenstein's arguments.

5.6.1 The anti-Felsenstein zone

Overcorrections for multiple substitutions often have the effect of increasingly overestimating the true distance, as the observed distance increases (e.g. Figure 2.6 but consider the i.r. distance to be the true distance). For most 4-taxon trees, application of a distance tree building method (e.g. neighbor joining) will choose the tree which minimises the sum of $\delta_{ij} + \delta_{kl}$, as discussed previously. In the Felsenstein-zone, parallel changes result in long edges attracting. In the anti-Felsenstein zone, accurate correction, or overcorrection, negates this effect and sees that the minimum of $\delta_{ij} + \delta_{kl}$ is always consistent in selecting the true tree, where *ij* vs *kl* are separated by an internal edge.

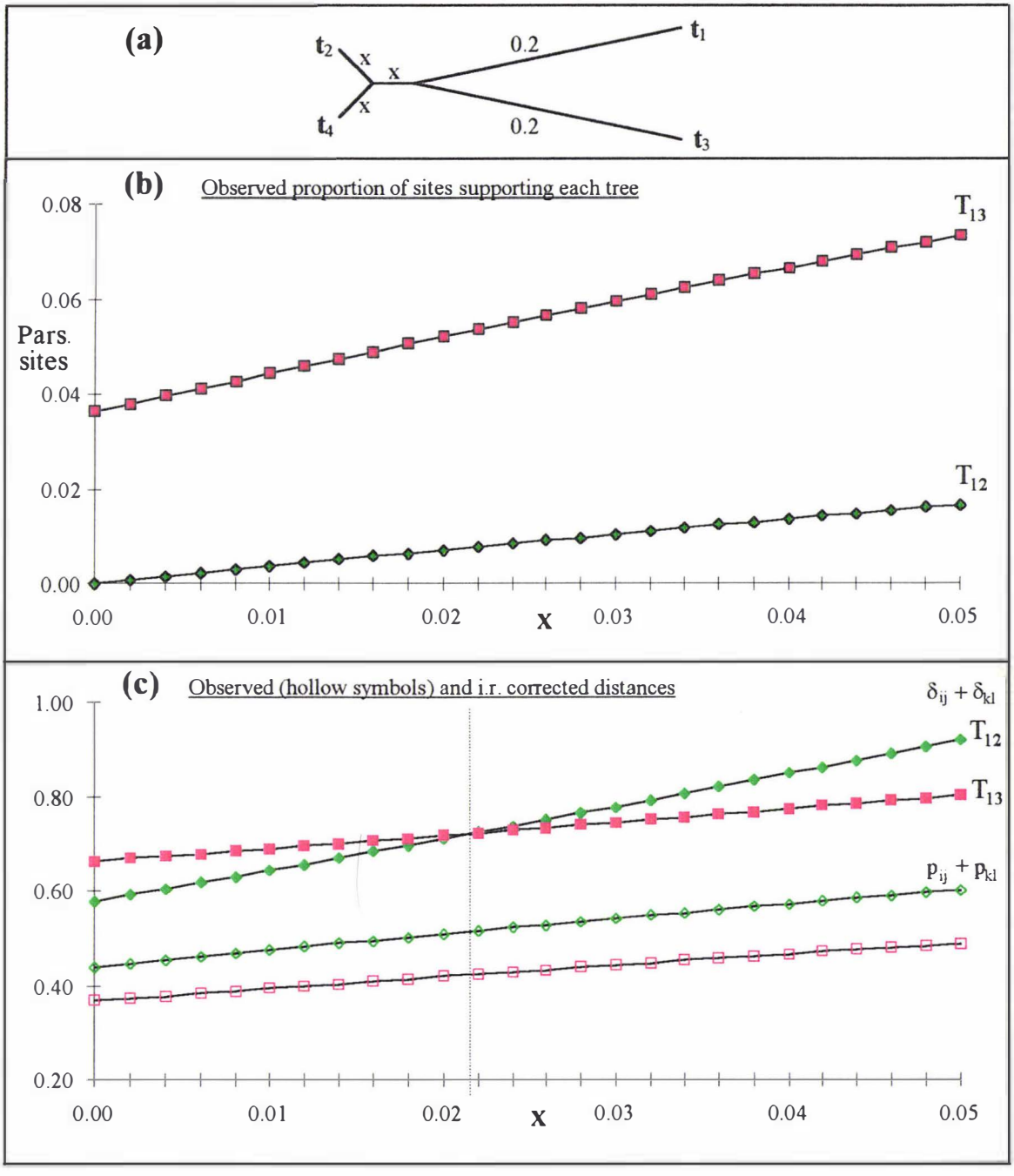
Overcorrection is a problem on a tree where two long edges are grouped together (here *ik*, or taxa 1 and 3 in figure 5.15a), with a short internal edge separating these from two short external edges (*jl*) (an anti-Felsenstein tree). Here overcorrection can have the effect of making the sum $\delta_{ij} + \delta_{kl}$ less than the sum which would be minimal with tree additive distances ($\delta_{ik} + \delta_{jl}$). This is because the largest distance on this tree is δ_{ik}, and it is this distance which is most severely overestimated (on the example of an anti-Felsenstein tree in figure 5.15a, $i = 1, k = 3, j = 2, l = 4$, for example). (Note: if the tree has bilateral symmetry, then $\delta_{ij} + \delta_{kl}$ equals the sum $d_{il} + d_{kj}$, but otherwise not necessarily so). The net result of overestimating the largest distances is 'long edges repel', or are pushed apart by the overcorrection.

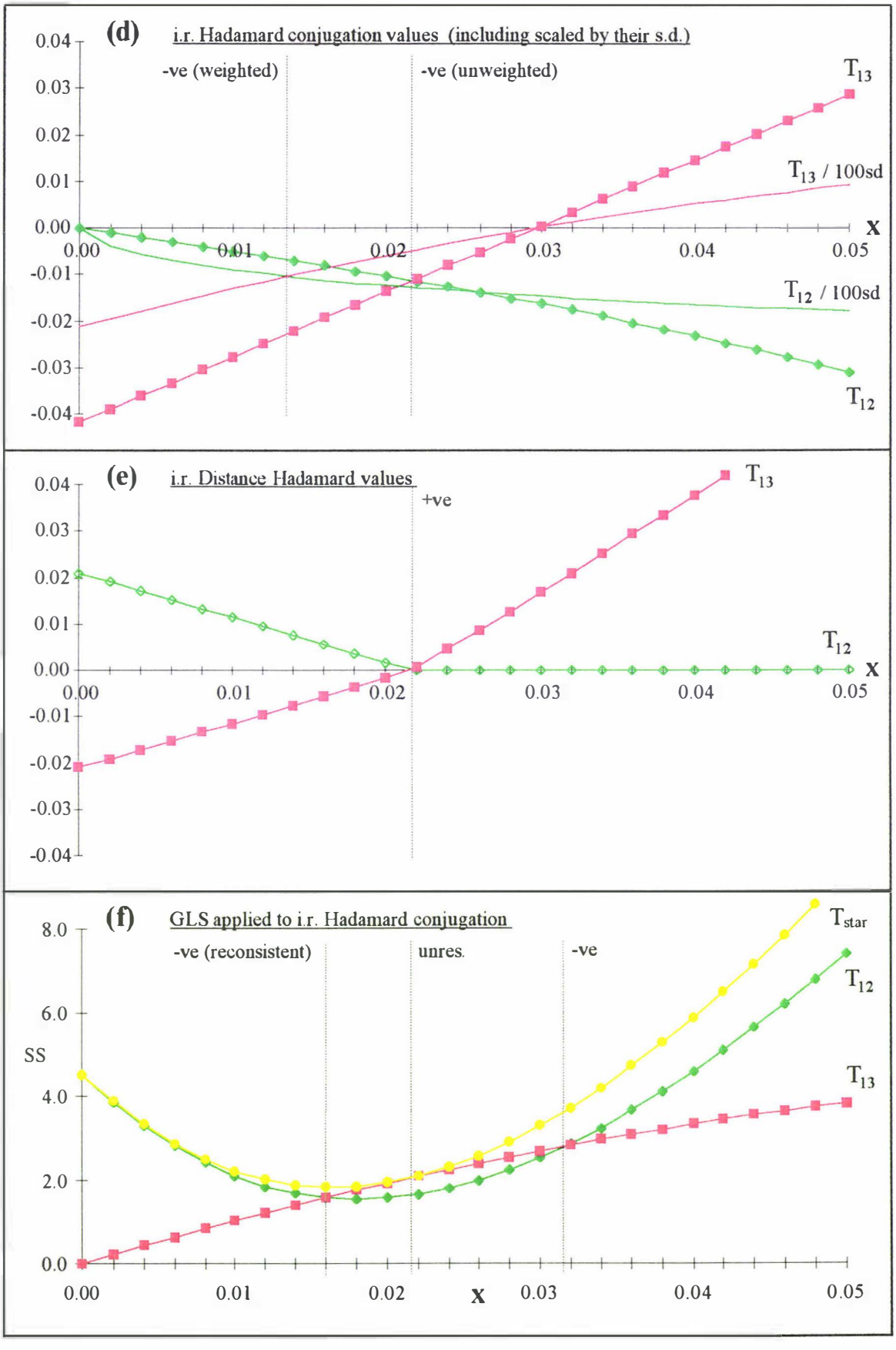
Let's now look at how much relative overestimation error is needed in order to get inconsistency in the anti-Felsenstein zone. Let er_1 be the overestimation of each distance δ_{ij} and δ_{kl}, less the overestimation of δ_{jl}, while er_2 is the overestimation of δ_{ik} (er standing for error term). If $er_2 - 2er_1 >$ {2 times length of the internal edge} (shown as *x* in figure 5.15a), then standard tree building methods like neighbor joining will be inconsistent in this region.

The previous result can be generalised to calculate how much error must occur to get inconsistency on any weighted 4-taxon tree, where the largest distance is between two taxa grouped together. Assuming that the true tree is T_{13}, inconsistency occurs when $d_{13} + d_{24} + er_2 > d_{12} + d_{34} + er_{d12} + er_{d34}$ (or replace right-hand side with $d_{14} + d_{23} + er_{d14} + er_{d23}$ if the sum of these

terms is less)(and where d_{ij} is the exact tree additive distance). Taking out edges in common to the tree additive distances we have inconsistency occurring if $er_2 > 2 \times$ internal edge weight + $er_{d12} + er_{d34}$, which rearranges to $er_2 - er_{d12} - er_{d34} > 2 \times$ internal edge weight. As a rough rule of thumb, if the maximum distance is less than 1 and the true distribution of rates is i.r., but we mistakenly assume a shape parameter of k =1, about as much potential error is introduced into distance estimation as if the converse were true (i.e. underestimation of distances, as figure 2.6 shows). In the case with Γ k = 1, if the c.v. of the distribution of rates across sites is increased again by a factor of two (so k = 0.25), this roughly doubles the amount of 'correction' added on to account for unseen changes. Thus if the true distribution of rates across sites is Γ k = 1, but we use a distance correction assuming Γ k = 0.25, this potentially leads to as much systematic error in tree selection, as parsimony applied to the observed sequences has. Thus we can state the converse of Felsenstein (1978) and Penny and Hendy (1989): With overcorrection of the data long edges repel.

(caption after figures; figures contain results for all subsections of section 5.6)


(a)
t2
x
x
x
t4
0.2
0.2
t1
t3
(b)
Observed proportion of sites supporting each tree
0.08
0.06
Pars. sites
0.04
0.02
0.00
T13
T12
0.00
0.01
0.02
x
0.03
0.04
0.05
(c)
Observed (hollow symbols) and i.r. corrected distances
1.00
0.80
0.60
0.40
0.20
δij + δkl
T12
T13
pij + pkl
0.00
0.01
0.02
x
0.03
0.04
0.05


(d)
i.r. Hadamard conjugation values (including scaled by their s.d.)
-ve (weighted)
-ve (unweighted)
T_{13}
T_{13} / 100sd
T_{12} / 100sd
T_{12}
x
0.04
0.03
0.02
0.01
0.00
-0.01
-0.02
-0.03
-0.04
0.00
0.01
0.02
0.03
0.04
0.05
(e)
i.r. Distance Hadamard values
+ve
T_{13}
T_{12}
x
(f)
GLS applied to i.r. Hadamard conjugation
-ve (reconsistent)
unres.
-ve
T_{star}
T_{12}
T_{13}
SS
8.0
6.0
4.0
2.0
0.0
x

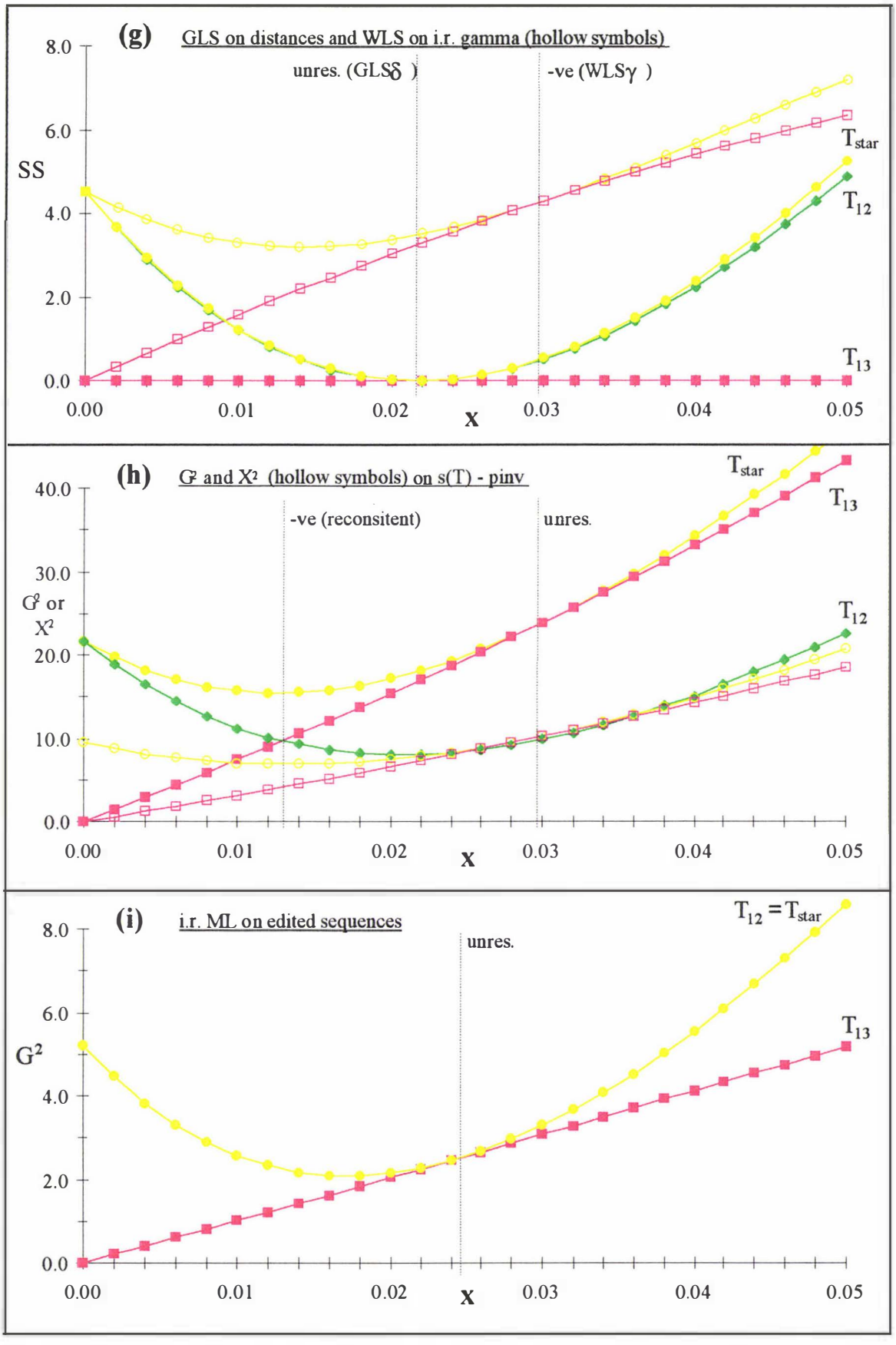


FIGURE **5.15a** Our model for the anti-Felsenstein zone of inconsistency consists of this weighted tree, while the mechanism of substitution is the 2-state equifrequency i.r. process. In this example inconsistency is achieved by assuming that 25% of all sites are invariant, when in fact no sites

are invariant. This effectively generates overcorrection for the number of multiple hits when i.r. transformations are applied. A very similar effect can be achieved assuming that rates across sites follow a Γ distribution with shape parameter $k = 1$. For the purpose of evaluating goodness-of-fit statistics, a nominal sequence length of 1000 has been used when calculating SS, G^2 etc. Results are shown in graphical from in figures **5.15b - 5.15h** and correspond to the same methods used in figures 5.11 - 5.13. The model tree is T_{13} marked with red squares, the incorrect tree T_{12} is marked with green diamonds, while the unresolved T_{star} is marked with yellow circles. Statistics pertaining to T_{12} are labeled, in all these examples the other incorrect tree, T_{14}, has identical values (due to tree and path length symmetries). The dotted vertical lines indicate the value of x at which a different tree is expected to be selected by one of the methods being shown (usually that method where lines touch or cross). If the line is marked with a "-ve" method becomes inconsistent if negative edge weights are allowed. If it is marked with a "unres." it means that unless negative edge weights are allowed the method will default to the star tree at that point. If marked with a "+ve" this is most serious, since it means inconsistency for the best tree with positive edge weights. For other details of labeling refer to figure 5.11.

5.6.2 Long edges repel effects with simple criteria applied to γ, $\gamma(d)$ and δ

Here we examine how robust tree selection methods are in the anti-Felsenstein zone. Our model tree is shown in figure 5.15a, and the overcorrection transformations is achieved with all methods by removing 25% of all sites, in the form of constant or unvaried sites, prior to application of what are then standard i.r. methods. As was done for the Felsenstein zone evaluations, the maximum pathlength through the tree is always less than 0.5 substitutions per site. Also considered is how well the various methods can detect that the data does not meet the models expectations, allowing a rejection of the model as potentially erroneous. All results are shown in one series of figures, labeled as figure 5.15a-i.

As figure 5.15b shows, parsimony or compatibility applied to the observed data will not become inconsistent on this tree where multiple substitutions on the long edges mimic (and augment) the reliable changes on the internal edge (here parsimony applied to the observed sequences is the same as neighbor joining, minimum evolution, ST method, and closest tree applied to the observed data). A similar trend holds for any number of taxa, as the exclusion of "uninformative characters" does not affect this combination of null transformation and tree selection criteria. (However, to describe when consistency occurs requires statements such as "no other multiple hits override the additional support that internal edge patterns get"). The same conclusion applies to tree selection from observed distances (not shown). Indeed, on this sort of tree, long edge attraction results in these methods recovering the true tree more often than would be expected if all the external edges where the same length.

However, as figure 5.15c shows, the addition of an i.r. logarithmic correction (which is really an i.r. + p_{inv} 0.25 correction) sees the minimum of the sum of distances switch from indicating the correct tree to an incorrect tree at $x = 0.0215$. At this value the neighbor joining algorithm becomes inconsistent, switching its preference to either of the two incorrect trees (they are both equal so which one is chosen depends on taxon order in the distance matrix). The estimated weight of the internal edge is positive, declining to zero at the transition point (0.0215), then again becomes positive (but incorrect) and this weight for an incorrect internal edge continues to

increase until $x = 0$. When x decreases to 0.009, the estimated length of the two short external edges goes to zero, becomes steadily more negative, as x goes to zero (but remain small in absolute size, results not shown).

It is informative to consider how the ordinary least squares upon distances (OLSδ) method performs (see Cavalli-Sforza and Edwards 1967, Swofford and Olsen 1990). For this model, and without any additional constraints, OLSδ will always find the true tree optimal, with the weight of the internal edge changing from positive to increasingly negative when x is less than 0.0215 (results not shown). For all values of x studied, the fit of distances to tree was perfect, when allowing negative edge lengths. The same result is obtained for the Fitch and Margoliash (1967) 'FM' weighted least squares distance method, when allowing negative edge lengths. If the constraint that all edge weights must remain positive is applied, then both OLSδ and FMδ become inconsistent at $x = 0.0215$ (the same value at which neighbor joining becomes inconsistent). For these two explicit sums of squares criteria, with the constraint that all edges must be ≥ 0, the SS is zero until $x = 0.0215$, but then increases as x goes to zero.

The neighbor joining method, with four taxa, is identical to the method of choosing the unconstrained OLS distance tree with the minimum sum of edge weights (not necessarily the best OLS fit)(Rzhetsky and Nei 1992a, verified here with numerical calculations). The weighted neighbor joining tree (i.e. with edge lengths) is identical to the favoured OLS tree until the critical value $x = 0.0215$. The weighted neighbor joining tree is then identical to either of the two incorrect trees with edge weights calculated by OLS until $x = 0.009$. For $x < 0.009$ the weighted neighbor joining tree is identical to either of the incorrect trees with edge weights calculated by OLS with negative edge weights allowed. Even amongst these closely related criteria, the tree being chosen varies when x is less than 0.0215. These criteria also give quite different indications that something may be amiss. These are in the form of rising sums of squares in the methods forced to have non-negative edge weights, versus negative internal edge weights, but perfect fit, when negative edge weights are allowed. We discuss the implications of these findings to the application of these tree building methods at the end of this section.

Figure 5.15d shows the i.r. Hadamard conjugation values which determine the optimal tree with the parsimony criterion. If the largest value relating to an internal edge is chosen as optimal, irrespective of whether it was negative, then the inconsistency point would occur for $x \leq 0.0215$. However, often we would not wish to pick any edge in our tree which did not have a positive value in the γ vector. Subsequently methods such as compatibility or closest tree applied to γ would not be inconsistent, but would default from resolving the trifurcation for values of $x \leq 0.0295$. So, when x is small the signs of entries in the Hadamard conjugation are cautioning us away from picking any resolved tree.

Also marked on figure 5.15d are the signals in γ pertaining to internal edges, divided by their standard deviation (as calculated by the analytic method of chapter 4, and assuming that the sequence length is 1000). The size of these entries indicates that the deviation towards negative values would frequently be detectable by statistical tests for the largest and smallest values of x

considered (at the nominal sequence length of 1000). That is, tests on γ should facilitate detection of lack of fit. The performance of the tree selection method WLSγ, which uses $\gamma_{s.e.}$, is considered later.

The values associated with potential internal edges calculated with the distance Hadamard (figure 5.15e) are not so cautionary. These signals are inferring the correct tree for x greater than 0.0215, but the zero values of the two non-tree signals provide no evidence that the model's assumptions are being violated. When x falls below 0.0215, standard tree selection criteria (e.g. compatibility, closest tree, parsimony) pick an incorrect tree with increasingly large internal edge weight as x goes to zero; an undesirable result. For $x < 0.215$, the size of the negative values associated with the other trees might serve as a warning of overcorrection of the data (although on average their statistical significance, in this example at least, would tend to be less than those in γ, since their expected values are less than half as large and negative, while variances will only be slightly smaller).

5.6.3 Performance of weighted least squares methods from γ and δ

Figure 5.15f shows GLS tree upon the data transformed by the Hadamard conjugation (GLSγ). If we reject selecting a tree with negative edge weights, then this method is consistent for larger x, and then defaults to picking the star tree. This is because the edge weight on T_{13} is positive until $x = 0.0215$, beyond this value T_{13} has a negative internal edge weight. As shown, the fit of T_{12} is only better than that of T_{star} because we have allowed this tree to have a negative internal edge length. Consequently, this method would only be inconsistent (i.e. pick an incorrect tree with very long sequences) if negative edge lengths are allowed. In figure 5.15f, the vertical dotted line marked -ve, shows the point at which inconsistency would occur if negative edge weights were allowed, while the line marked "+ve" shows the point at which an unresolved tree would be chosen if the internal edge must be non-negative. Notice how inconsistency when allowing negative edge weights can only occur in an interval of possible x values; in this case x between 0.016 and 0.0316, because for smaller values of x T_{13} again becomes the best fitting tree (in figure 5.15 this point is marked as 'reconsistent'). Overall, this criterion, by way of the sum of squares, is providing a clear sign that the data is not fitting the model (remembering of course to add the additional lack of fit expected due to sampling error). Here the true distribution of the SS asymptotically becomes a non-central chi-square (Stuart and Ord 1990, p 867).

The effect of using only distances, rather than γ, is shown in the next figure (figure 5.15g) as the lines with the solid symbols. The performance of GLSδ is very similar to that of OLSδ. If negative edge weights are allowed, GLSδ always picks T_{13} with no indication that the model is violated, i.e. this tree always has a perfect fit to the pairwise distance data. If all edge weights must be positive, then the method will default to the star tree when x is less than 0.0215. Consistent with what was observed earlier in the Felsenstein zone, indications of lack of fit of data by expected residual SS are not as pronounced with GLSδ as they are for GLSγ (compare y-axis values on figure 5.15f with those on 5.15g). The reason that distance methods can fit this data perfectly is due to there being fewer degrees of freedom in distance data compared to

sequence data. It is not expected that this property of perfect fit will generally extend to examples with more than four taxa.

The residual SS by WLS on γ is also shown in figure 5.15g as the lines with hollow symbols. These calculations show similar indications of lack of fit to GLSγ, except that the SS are inflated due to WLSγ not taking into account correlations between entries in $\hat{\gamma}$. If we specify that all internal edge weights must be positive, then WLSγ will default to selecting the star tree when x is less than 0.0295, and so will not be inconsistent. If determined to pick a resolved tree, negative edge length or not, then tree selection from $\hat{\gamma}$ after weighting by the inverse of estimated standard errors, will give some robustness (compared to not using weighting). In this case the method of WLSγ will become inconsistent when x becomes less than 0.0135, whereas without weighting x need only be less than 0.0215 for inconsistency to occur (both points can be seen in figure 5.15d). Here also, the behaviour of WLSγ is unlike that of GLSγ when negative edge weights are allowed (compare figures 5.15f with 5.15g). This is the first time we have noticed a distinct difference in their properties.

5.6.4 "Goodness-of-fit criteria" measured on the observed sequences

The last methods to be evaluated here are the methods which measure fit with well known multinomial statistics applied at the **s** level. Figure 4.15h shows the fit of the observed to predicted data (measured by G^2 and X^2) using as tree edge weights the values in $\hat{\gamma}$. As long as internal edge lengths must be positive, these methods default to the star tree at the same value of x (namely 0.0295) that the standard tree selection methods applied to $\hat{\gamma}$ do (see figure 5.15d). For both $G^2\hat{\gamma}(T)$ and $X^2\hat{\gamma}(T)$, if negative edge weights are allowed, then tree selection by either of these criteria will be <u>inconsistent</u> for all values of x, except when x is very small (here x less than 0.0125)!

The previous results are at first surprising, but still understandable. In the Felsenstein zone, the parsimony pattern corresponding to the true tree showed a lot of leverage on the overall likelihood due to both the form of **s**(T) taken from $\hat{\gamma}$, and the inherent sensitivities of these statistics. (For example, to give more weight to common patterns which are expected to be rare, versus rare patterns which are expected to be common, and by leverage we mean these relatively rare patterns plays a disproportionately large role in determining overall fit). In the anti-Felsenstein zone a similar thing is happening; when negative edge weights are allowed **s**(T) can 'explain' why the two incorrect signals are so rare, and it is these patterns which are then offering the most leverage (by way of reducing the lack of fit). As long as negative edge weights are not allowed these methods do not become inconsistent. Overall, these methods appear to be showing good evidence of the lack of fit of data to model, although as yet the exact sampling distribution of these fit statistics is unknown. In a reversal of what we saw in the Felsenstein zone, it is the G^2 statistic which is more powerful than the X^2 statistic at detecting deviations of the data from the models expectation.

The presence of negative edge weights in calculating likelihoods is a newly observed phenomena. At present it is uncertain if the likelihoods of the data patterns calculated in this way, using Hadamard conjugations, can be considered mathematically 'correct' (the real world difficulties of giving a concrete interpretation to negative edge lengths aside). One attitude would be that negative edge weights are never allowed, so taking this view, these methods would not become inconsistent (for this model).

Lastly, the ML tree selection method is evaluated (fig. 5.15i). As long as negative edge weights are forbidden, then this method is consistent for $x > 0.0243$; for x below this value, the method defaults to the star tree. For most of its range the method is also giving a good indication of lack of fit of data to model, via the G^2 statistic. The method of minimum X^2 was also evaluated (data not shown). The point of defaulting to the star tree was fractionally higher than that with the G^2 method (just 0.0002 being the difference, which is so little that if marking the point on this figure, the lines would almost touch). Generally the minimum X^2 statistic was 10-20% less than the size of minimum G^2, indicating a slight reversal of the relative sensitivity of these statistics with respect to their performance in the Felsenstein zone. Interestingly, as noted earlier, it seems to be possible to calculate the likelihood of data with relatively small negative internal edge weights using the Hadamard conjugation. With larger negative edge weights in $\gamma(T)$, entries in $\mathbf{s}(T)$ can become negative, which makes it impossible to interpret these $\mathbf{s}$ values as probabilities. This, however, presently appears to be of little useful application, as allowing negative edge weights can mislead these tree selection criteria.

5.6.5 Summary of tree selection in the anti-Felsenstein zone and its implications

With overcorrection of the data, all these methods overestimated the length of external edges (especially the longest edges), but underestimated the length of the true trees internal edge. Such findings are intuitively reasonable (since they are explaining part of the support for the internal edges as multiple hits) and are the exact opposite of what happens with insufficient correction for multiple substitutions, as in the Felsenstein zone (section 5.5).

It seems fair to say that all the more sophisticated statistical tree selection methods did quite well in these circumstances, not selecting an incorrect tree unless negative edge weights were allowed, and generally giving a moderate indication that the data was not fitting the model (both by the fit statistic and the presence of negative internal edge lengths). The Hadamard conjugation (but not so much the distance Hadamard) also gave a clear picture that overcorrection of the data seemed to be occurring.

In these simple studies, of all the transformations and criteria evaluated, only two would consistently select an incorrect tree in the anti-Felsenstein zone, given the reasonable constraint of not allowing negative internal edge lengths. These were the neighbor joining algorithm (here equal to the minimum evolution criteria), and the distance Hadamard method with standard tree selection methods (both of course applied to overcorrected data). This finding is ironic, in that both sets of authors identified at the beginning of this section as advocating (or using) transformations which could overcorrect (Jin and Nei 1990, Lockhart *et al.* 1994), also used

neighbor joining and / or the distance Hadamard. In contrast, the approach we have emphasised in this thesis (especially in chapter 3), aims to minimise the anti-Felsenstein problem by estimating either a best-fitting proportion of sites to remove or the optimal shape parameter for a given distribution, and not relying upon a fixed quantity.

Even with optimisation of fit of model to data, it is possible to assume that too many sites are invariant, or assume a shape parameter which is too extreme. After selecting a tree by neighbor joining or the Fitch program (Phylip, Felsenstein 1993), searching in its vicinity with unconstrained OLS could be interesting. If this turns up an optimal tree with negative edge weights, there may be cause for suspicion. It could then be checked what proportion of pairs of distances which could resolve this edge agreed with this resolution. We expect this should offer an interesting test of the reliability of tree selection when rates vary across sites, and also when the process and rate of substitution varies from edge to edge. Of course caution must still be advised when we are analysing anciently diverged molecules for which a covarion type model may (must if the molecules are still informative after about 500 million years) be the major force molding the process of evolution.

A salient point which comes out of these studies, is that in reality the interesting phylogenetic problems generally are trees with long external edges and short internal ones, and presently we do not know much about the real process of evolution (e.g. covarion models, structural-functional shifts). It could turn out that under at least some real processes could be yielding data for which i.i.d. transformations would overestimate the larger distances, resulting in a real danger of anti-Felsenstein effects. Such an effect could be duplicated in many different genes, so that congruence would not be a reliable indicator of consistency. At present we simply do not know the scope of this problem. It could be a very hard problem to detect, as it could occur in some parts of a phylogeny but not others, if the covarion process of evolution had itself been substantially non-homogeneous. The conclusion must be clear: very deep, ancient divergences, with long external edges cannot reliably be resolved with our current model based methods (e.g. see the quote at the beginning of chapter 3). It is for this reason that I believe model based phylogenetics cannot presently completely displace a more traditional type of systematics, based on identifying very slowly evolving and outwardly reliable characters. It is these most conservative characters which one must hope corroborate the findings of the more sophisticated model based methods. Our counting of "perfect characters" in deeply diverged 16S-like rRNA in chapter 3 is an example of the application of just this sort of philosophy (as is congruence with a character such as 5.8S fission from the LSU rRNA).

Lastly, it is important to appreciate that understanding the anti-Felsenstein zone has implications for simulation studies. When evaluating a tree selection method with a bias towards increasing overestimation as the observed distance increases then, if a tree is in something like the Felsenstein zone, such a method can do better than an unbiased method (since overestimating the largest distance will bias the results towards selecting a long edges separated tree). To detect this "false statistical efficiency" of tree selection, it is necessary to compare methods also in the anti-Felsenstein zone where such a bias will work against the reliability a tree building procedure

(and also on a tree with equal edge weights, but a short internal edge, where overall error from tree additive distances diminishes the efficiency of a tree selection method). Logically, all we know about a tree with a very short internal edge is that it could group the external edges in any order.

Desirable properties of a tree selection method should be that it will select the true tree with high statistical efficiency, irrespective of what the true tree is. Another desirable property is that it should not be biased in this estimation (e.g. see Kuhner and Felsenstein 1994 for a discussion and evaluation of bias in tree selection). This must mean that it will pick the correct tree about equally frequently, irrespective of the arrangement of edge lengths on the 'unknown' true tree. This last point is important, since it usually makes little sense to use a highly biased method on an unknown problem. Doing this, puts an onus on the researcher to evaluate systematic error when presenting a statistical summary of their results (making the analysis unnecessarily complicated).

Recently I applied this prediction to some simulations run with Dr David Swofford. The program was that used in Hills *et al.* (1994), the tree was a Felsenstein tree, and the mechanism of evolution was i.r. stationary 2P Kimura. Under this model the LogDet in combination with neighbor joining was recovering the true tree more often than the Kimura 2P distance also in combination with neighbor joining. I had doubts about the generality of these results from the theory described in this section, so the program was modified to allow anti-Felsenstein trees also. When the simulations were rerun in the anti-Felsenstein zone the tree building procedure using the Kimura 2P transformation did better than when using the LogDet (and the difference was sufficient that the average reliability of the Kimura 2P transformation over both Felsenstein and anti-Felsenstein zones was now slightly better than that of the LogDet).

This observation makes the point that the type of tree being used in a simulation can result in a biased picture of the relative efficiencies of different methods. In order to get a true picture of the reliability of a tree building procedure (even under our vastly simplified models) one needs to include in a study trees with edge lengths that both narrowly separated long edges alongside short edges and narrowly grouped long edges along side short edges (since in application to real sequences we will not know which is correct). (It is also interesting to note that in biology anti-Felsenstein situations should be more wide spread than Felsenstein trees since they require fewer changes of the biological rate of substitution to occur, e.g. one rate change in figure 5.15a versus two rate changes for the tree in figure 5.11a). It is also important to consider the reliability of the resolution of trees which are near clock-like, but difficult to resolve due to short internal edges.

All in all these considerations suggest that recent large scale simulations based solely on the Felsenstein zone (e.g. Huelsenbeck and Hillis 1993) need to be expanded considering also the anti-Felsenstein zone, all equal external edge lengths, and tree selection bias, if they are to be cited as representative of even four taxa. Simulations should attempt to unravel factors of bias due to topology plus edge lengths (we call this tree shape), and at least attempt to imitate reality in critical ways if they are to be of use guides to practical phylogenetics. If they are to test or evaluate certain conjectures, then they should be aimed at this. When they fall in between these

two possibilities, their value rapidly diminishes. The results presented here should help to give a better understanding of the factors which should be taken into account by future simulation studies, and stimulate interest in finding other combinations of edge lengths which may give unpredictable results. One such possibility (D. Penny pers comm.) is a long edges attract tree, which requires only one lineage to deviate from clock-like in order to obtain inconsistency, with two trees being shorter than the true tree (e.g. a tree like (A: 0.3, B: 0.05):0.02,(C: 0.2, D: 0.2), in the notation of Felsenstein 1993).

5.7. INCONSISTENCY OF ML IN THE HENDY-PENNY ZONE

Sections 5.5 and 5.6 show that maximum likelihood and every other method we tried could be inconsistent (in selecting the correct unweighted tree) when the effect of unequal rates across sites was not adequately accounted for. We now show that another area where parsimony applied to observed sequences is known to be inconsistent, the so called 'Penny-Hendy zone' (Hendy and Penny 1989), is indeed also a zone of inconsistency for ML. A Hendy-Penny tree is a five taxon tree (they also suggested a six taxon variant) which obeys a molecular clock, but has short internal edge lengths relative to tip lengths, and a more distant outgroup taxon (an example of a Hendy-Penny tree is shown in figure 5.4). Also considered is the consistency and robustness of parsimony applied to i.r. γ for data generated by this new model (the 'corrected parsimony' of Steel *et al.* 1993b).

5.7.1 The Hendy-Penny zone

In the Hendy-Penny type of tree, as rates of change increase, there are a substantial number of parallel and convergent substitutions especially between the outgroup and ingroup taxa (as an example of a Hendy-Penny tree see figure 5.4, the tree on the right ignoring the dotted edge). As the internal edge length decreases, there comes a point where these parallel changes become more frequent than the site patterns supporting the internal edges. At this point four trees become better (shorter by parsimony) than the true tree. These are trees where the outgroup edge joins either of the four ingroup external edges, to give four different trees (thus breaking up just one of the two ingroup clusters). If the internal edge becomes shorter, or the number of convergent and parallel changes increase further (either by increasing external edge lengths, or by the model incurring more multiple hits) then eight more trees can become shorter than the true tree. How this happens, is in a sense, a progression how the four wrong trees were generated. On each of these four trees, either of the other two external edges still correctly grouped together, are attracted to the long outgroup sequence, and may join with it, generating 8 (four times two) more incorrect trees shorter than the true tree. The only two trees guaranteed to be longer (by the unweighted parsimony criterion applied to the observed sequence patterns) than the true tree are the two interchange trees, where we have the groupings ((A,C),(B,D),E) or ((A,D),(B,C),E) (relative to the tree in figure 5.4). (The description here is for 2-state data, a description for 4-state data is similar).

The model we will use to further examine properties of the Hendy-Penny zone is the 2-state Poisson model, using the tree described in figure 5.16 The x-axis variable is the length of the external edge leading to the outgroup (marked "E length", and running from zero up to 5 substitutions per site, counting multiple substitutions at a site). The green line shows how much longer any one of the four wrong trees is than the true tree, when lengths are measured on the correct $\gamma(T)$ (the exact number of substitutions per site, or exactly 'corrected parsimony'). Clearly the distance to the outgroup does not affect this difference, and the increased length of these incorrect trees is exactly the number of extra changes required to break one of the two internal edges (each of length 0.01).

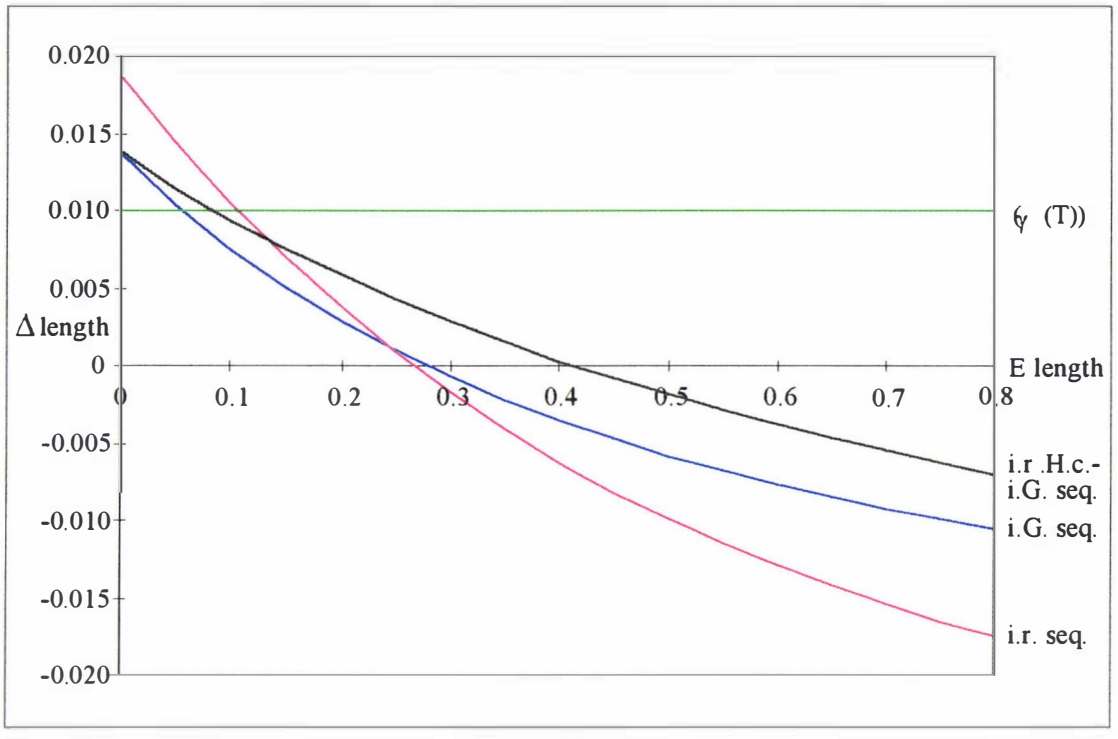


FIGURE 5.16 Inconsistency of parsimony applied to the observed sequences and parsimony applied to i.r. Hadamard conjugation transformed sequences, under a molecular clock. The model is 2-state Poisson, with rates across sites following either an i.r. or an inverse-Gaussian (i.G.) distribution (shape parameter = 1), with the true weighted tree being (((A: 0.2, B: 0.2), 0.01), ((C: 0.2, D: 0.2), 0.01), E: X), where X is the length, counting multiple substitutions, of the edge leading to the outgroup (x-axis of figures). The Y-axis measures how much longer (per site) the any of the four wrong trees are relative to the true tree (when this number becomes negative inconsistency occurs). The labels are light green line, parsimony applied to $\gamma(T)$, red line, parsimony applied to the observed sequences when rates across sites are identical, blue line, parsimony applied to the observed sequences generated when the distribution of rates across sites is inverse Gaussian, shape parameter equals one. Lastly the dark green line is parsimony applied to the sequence pattern probabilities generated under the i.G. model, but then transformed by the i.r. Hadamard conjugation. Here, identical results will be found with compatibility or closest tree as the tree selection criterion.

The red line of figure 5.16 shows the difference in length of a wrong tree to the true tree when measured on the observed i.r. evolved sequences. In this example the difference in tree length is simple to calculate directly from s(T). Firstly, the constant and singleton patterns are ignored. The 'parsimony informative sites', either fit a tree and require one change, or if they do not fit the tree, they require two changes. The only patterns which have a different score on the true, versus one of the four wrong trees, is the signal for an internal edge, versus the signal for the false pattern grouping the outgroup with one of the ingroup sequences. Consequently, the difference in tree length measured per site on the observed sequence as shown in figure 5.16 is just the difference of two patterns in s(T), each corresponding to an internal edge of the two trees. As figure 5.16 shows, the true tree starts out distinctly shortest (by about 18 steps if the sequences were 1000 long), but when the edge to the outgroup exceeds 0.27, inconsistency occurs, and gets worse until an asymptote is reached when the outgroup sequence becomes random with respect to the ingroup sequences. At this point the difference in length between the (continued)

true and false trees is -0.02652 per site (i.e. for a sequence of length 1000 the true tree is expected to the 26.5 steps longer than any of the four wrong trees).

5.7.2 The Hendy-Penny zone with unequal rates of change across sites

We now extend the model to allow sites to evolve at unequal rates, but all other variables remaining the same as previously. The rates to follow an inverse Gaussian distribution, with shape parameter equal to one (so c.v. = 1 also). The difference in parsimony length measured on $\gamma(T)$ (the light green line) remains the same, but the difference in parsimony lengths of the true and wrong tree measured on the observed sequences has changed, as expected, and is shown by the blue line. Interestingly, two indicators of the degree of inconsistency have decreased: the edge leading to the outgroup sequence must be slightly longer for inconsistency to occur, and the asymptote for the blue line is less than that of the red line being -0.02069. The dark green line shows parsimony applied to the i.r. Hadamard conjugation transformation of observed sequences evolving with an inverse Gaussian distribution of rates across sites. This tree selection procedure has improved robustness, but it too becomes inconsistent. Indeed the asymptotic value is now larger than with the other methods (of scale to the right); when the outgroup sequence is random, parsimony applied to the i.r. Hadamard conjugation infers that the true tree is longer than the four incorrect trees by -0.04846 per site (or an expected 48.5 substitutions if the sequences were 1000 long). Clearly, for this sort of reason it was desirable to develop the Hadamard conjugations of chapter 2, in order to counter this effect.

5.7.3 Showing ML to be inconsistent in the Hendy-Penny zone with URAS

If maximum likelihood tree selection based on the i.r. 2-state Poisson model, is applied to infinitely long sequences based on the same model we recover the correct weighted tree in all instances (except when the outgroup sequence is random, and so can join the tree of the ingroup sequences anywhere with equal likelihood). However, if we apply this standard i.r. ML method to the sequences having evolved with unequal rates across sites inconsistency can occur. We show this in figure 5.17a, where i.r. ML is applied to exactly the same trees used to generate figure 5.16. The blue line is the G^2 or likelihood ratio statistic of the true tree, while the purple line is the G^2 statistic of any of the four wrong trees. After a slight bump in the curve with short outgroup sequences, the G^2 statistic begins a slow decline which asymptotes at 6.13 as the outgroup sequence goes to random (with respect to the ingroup sequences). For sequences of about 1000 long, a goodness-of-fit of 12.59 or worse would lead to rejection of the model (at the 0.05 level) (ignoring the effect of sampling error on the G^2 statistic).

(figure next)

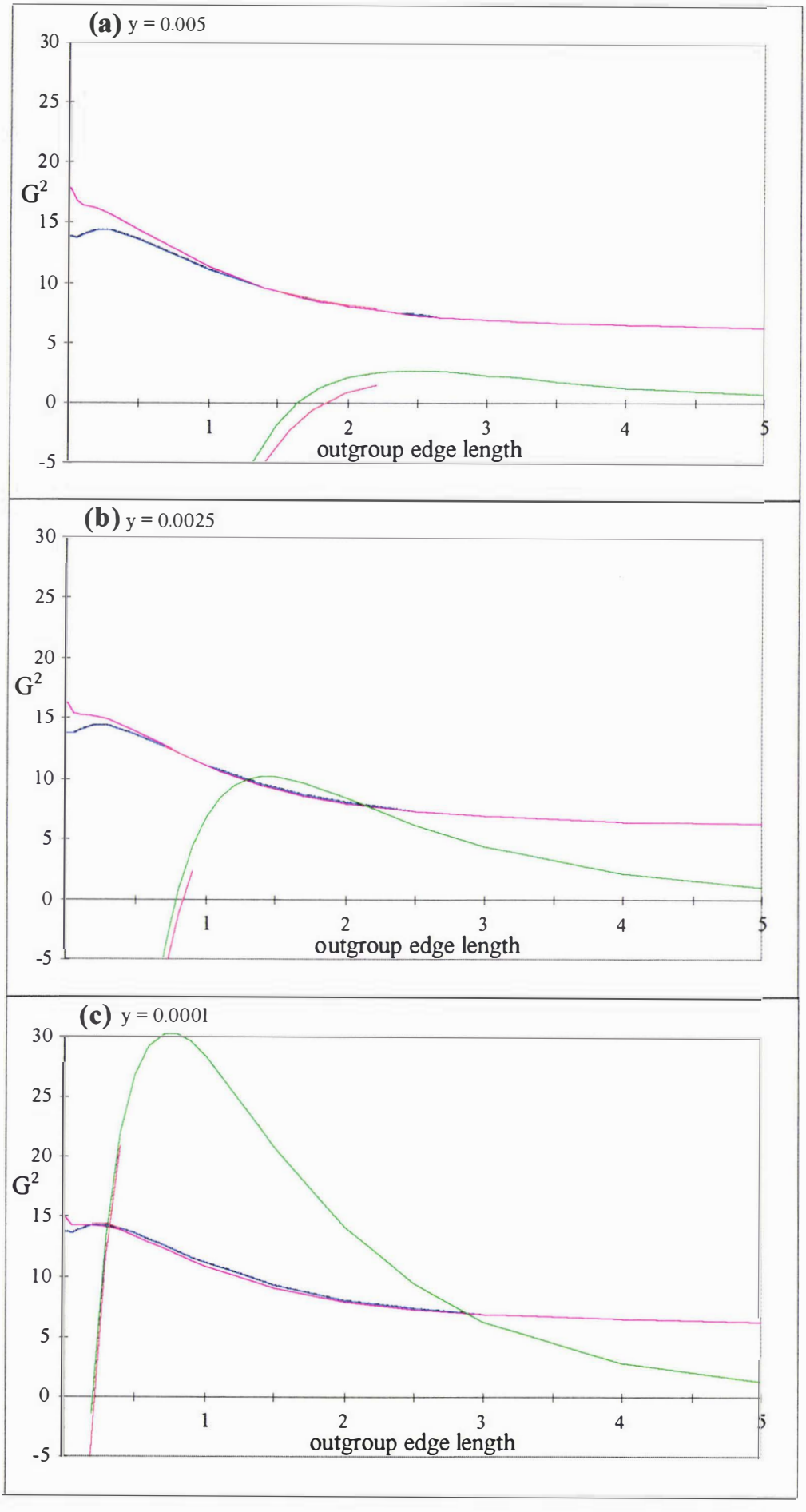

(a) y = 0.005
30
25
20
G2
15
10
5
0
-5
1
2
3
4
5
outgroup edge length
(b) y = 0.0025
30
25
20
G2
15
10
5
0
-5
1
2
3
4
5
outgroup edge length
(c) y = 0.0001
30
25
20
G2
15
10
5
0
-5
1
2
3
4
5
outgroup edge length

FIGURE **5.17a-c** Inconsistency of ML in the Hendy-Penny zone. The sequences evolved under the 2-state Poisson model, with rates across sites following an inverse-Gaussian distribution, with the true tree being ((A:0.2, B:0.2): Y, (D:0.2, E:0.2):Y, E:X), where Y is the internal edge length given at the top of each figure, while X is the total length of the edge leading to the outgroup (x-axis of figures). The blue lines are the G^2 statistic of the true tree (using a nominal sequence length of 1000), while the purple lines are G^2 of the 4 incorrect trees where the outgroup associates specifically with one of the ingroup lineages, and the second pair of ingroup taxa remain correctly grouped. Clearly these values are very similar, so we have shown their difference (true tree G^2 - false tree G^2) multiplied by 100 as the green lines (when the green line takes on positive values, the wrong tree is chosen by maximum likelihood). Fitting by minimum X^2 gave very similar performance to G^2, although with slightly better robustness (we have the critical values for minimum X^2 as the truncated red lines).

Thus indication of a significantly bad fit of data to model would be near certain if the outgroup edge was short, but less likely as the outgroup got longer and inconsistency of tree selection begins occurring. This is a further indication that goodness-of-fit of model to all the data does reliably answer the biologists most pressing question, "can I trust the results in terms of the unweighted tree I obtain?" The fit of a wrong tree to this data is generally very similar to that of the true tree, and the fit of these four wrong trees converges towards the same asymptote as the true tree (which just means the ingroup subtree is the same, and the outgroup sequence is random so it can be located anywhere on this subtree with equal likelihood).

We have found that the actual difference in likelihood between the true and the wrong trees tends to be close in the Hendy-Penny zone when using moderate amounts of sequence divergence (certainly closer than the difference parsimony length measured on observed sequences). To more clearly show when inconsistency occurs, we have plotted one hundred times the difference in the log likelihoods of the false tree as the green line in figure 5.17a (if this line is positive, either of the four wrong trees has higher likelihood than the true tree). With $y = 0.005$, inconsistency occurs when the length of the edge leading to the outgroup exceeds approximately 1.6. Thus inconsistency of i.r. ML does indeed occur in this zone, but it is quite limited in scope under this weighted tree (the internal edges of this model tree are long enough that it almost does not occur). Once the edge leading to the outgroup exceeds 1 substitution per site, a likely outcome when analysing samples of 1000 sites would be a statistical tie between the true and at least one of the wrong trees.

The short red lines mark the fit of different trees optimised by the X^2 goodness-of-fit criterion. Clearly the fit is very similar to that inferred by the G^2 statistic. Interestingly, and like the earlier example in the Felsenstein zone, the X^2 tree selection criterion shows slightly more robustness than the likelihood criterion (although by a much decreased amount). The difference in X^2 between trees tended to be less than that by the G^2 criterion, and there was never as big a gap favouring the wrong tree.

Decreasing the internal edge length by 1/2, we arrive at the results shown in figure 5.17b. Inconsistency is occurring more readily, and more severely, although the difference in likelihood is still small in the zone of inconsistency. The minimum X^2 criterion again performs in a very similar manner to maximum likelihood, and again has slightly better robustness. The third figure

(5.17c) shows the result of calculations when decreasing the internal edge length to just 0.0001 changes per site. Now, the outgroup does not need to be very distant before inconsistency occurs (and occurs even more readily with parsimony applied to the observed sequences or an i.r. Hadamard conjugation of the observed sequences). Finally the difference in likelihood between the two trees in the zone of inconsistency is becoming just large enough that with the longer sequence lengths used in studies these days we might expect to statistically reject the true tree most of the time (with sequences 10,000 base pairs long, it is expected the wrong tree will be about 2.0 to 3.0 G^2 units worse, making rejection of the true tree using a statistic like the Akaike information criterion quite likely, see Miller 1990).

In conclusion, the Hendy-Penny zone is indeed a zone in which i.r. ML can become inconsistent when it does not take account of unequal rates across sites. It appears, however, that edge lengths need to be considerably longer than in the Felsenstein zone for this to occur. Counterbalancing this fact, is that with real data we may encounter Hendy-Penny trees more often than Felsenstein trees due to the quasi-clock like evolution of many molecules. At first glance, these results suggest that in the Hendy-Penny zone, the problem of inconsistency is much less for i.r. likelihood than for parsimony applied to observed or i.r. transformed sequences. This is a pleasing result, but it may be too soon to call it, as we have only explored one small corner of the parameter space (which can be represented as a three dimensional cube, with X, Y and Z the distinct edge lengths of the Hendy-Penny tree as its axes). Further, keeping an eye on reality, we should also consider in detail a fusion of the Hendy-Penny and Felsenstein zones. This is a tree like ((A:0.2, B:0.3):0.005, (C:0.2, D:0.3):0.05, E:X), where rate inequality is appearing amongst the ingroup taxa (here both B and D, perhaps more likely just one of the taxa). Situations like this are reminiscent of the problems resolving mammalian divergences. Here, the evolution of taxa is quasi-clock-like overall, with distant outgroups (the marsupials and monotremes), and the possibility of some faster evolving ingroups (rodents and bats, being suggested examples).

This sections results do look promising for ML, but when making analyses of real data, it is always best to remember: Maximum likelihood, like any tree selection criterion, can only be trusted when we understand the relationship between the real data and the model.

5. 8 STATISTICAL EFFICIENCY OF TREE SELECTION ON S, $\gamma(S)$ AND $\gamma(D)$

This section addresses the question of how reliable and statistically efficient tree selection from $\hat{\gamma}$ is. The proportion of times a tree selection procedure (transformation followed by selection criterion and search strategy) selects the tree that generated the data is the (statistical) efficiency of that estimator. Some authors have called this statistic the power of the method; we prefer to use power in the more formal statistical sense of being the ability of a test to reject the null model when it is untrue (such statistics are described in chapter 6). In this section we will refer directly to the

property of interest, the ability to pick the generating or correct tree form a random sample, so as to avoid confusion of the two dominant meanings of efficiency, computational and statistical. In particular we consider what effect the mean vectors and sampling distribution of $\hat{\mathbf{s}}$, (the observed data), $\hat{\gamma}(d)$ (the spectra estimated from just pairwise distances) and $\hat{\gamma}(s)$ (the result of a Hadamard conjugation) are having on tree selection, under a specific model. In order to make comparisons fair the same tree selection criterion (here compatibility) is applied, but each sample is transformed in three different ways (the null transformation giving the observed data, $\hat{\gamma}(d)$ and $\hat{\gamma}(s)$). In our evaluations, we also ask the question of what overall rate of substitution maximises the ability of a tree selection method to recover the true tree, then attempt to relate this back to fundamental properties of the procedures. The hypothesis form section 4.5.2, that the peak in tree recovery coincides with the values of path length correction for which the signal to noise ratio is minimised, is tested. In the last part of this section, how modifications to the model (different edge weights, a distribution of rates across sites) affect conclusions, are evaluated.

5.8.1 A six taxon tree model to evaluate tree selection procedures

The model tree for this section is like that shown in figure 4.13, a six taxon "caterpillar" tree with all external and internal edge weights equal, except for the most internal edge which has weight 1/10 that of the other edges (edge weights are measured in the expected number of substitutions per site). Sequences are 2 character states, and evolve by the standard Poisson model (Hendy and Penny 1993). The tree selection criterion in all cases was the compatibility criteria (the largest sum of weighted compatible patterns). Here, compatibility is expected to give nearly identical results to parsimony or closest tree, since there are just six taxa, two states, and usually only one edge unlikely to be resolved correctly. In the running of these simulations, each weighted model tree produced an s(T) vector. From this 1,000 sequence patterns were randomly sampled, and arranged into an $\hat{\mathbf{s}}$ vector (equivalent under the model to a random sequence of the same length). Each time an optimal tree was selected from $\hat{\mathbf{s}}$, statistics of interest to our evaluations were recorded (namely the tree and sums of absolute deviations for all patterns (except the constant sites), and the sum of patterns not in the tree). When selecting a tree from $\hat{\gamma}(d)$ or $\hat{\gamma}(s)$, each $\hat{\mathbf{s}}$ sample was Hadamard transformed to the appropriate vector. Each point in the following graphs represent doing this 40,000 times. We used this large sample size so that the standard errors (which are binomial in nature) for the results shown in the following graphs are typically no larger than the third significant place, and of negligible importance to any of our conclusions. These simulations were run using the program Hadtree (Penny *et al.* 1993) after D. Penny made modifications to keep the residual statistics of interest.

5.8.2 Features of tree selection from $\hat{\mathbf{s}}$

Figure 5.18a shows the results of tree selection using the compatibility criterion applied to $\hat{\mathbf{s}}$ with increasing rates of change across the whole tree (the weight for all edges but the short edge 7, which is 1/10 of this value). As mentioned earlier in section 4.7.2, the only edge likely to be got wrong is the short internal edge, and the only alternatives which are compatible with the two well supported internal edges are taxa 1, 2, and 4 together (indexed as 11) or taxa 3 and 4 together (12).

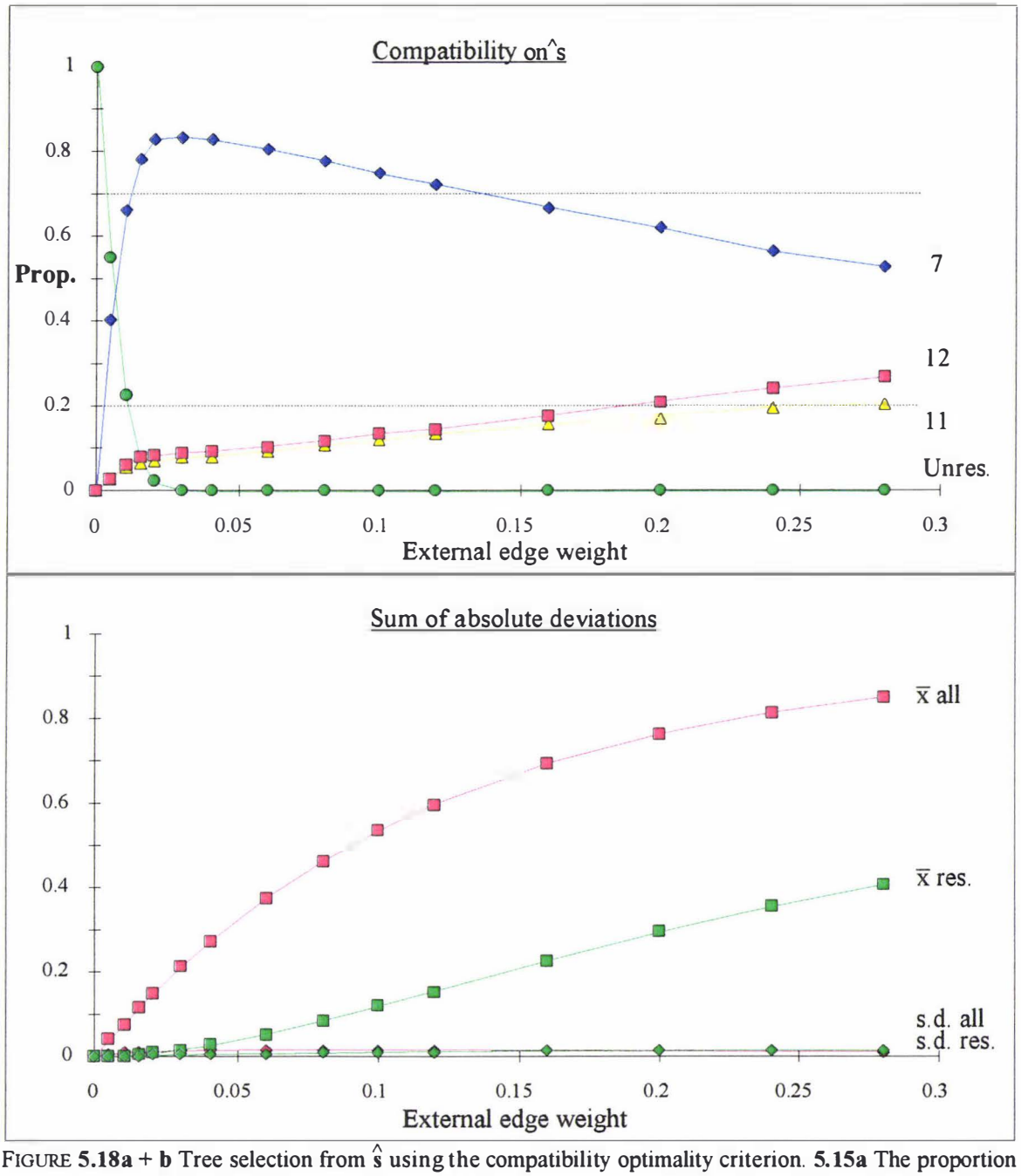


FIGURE **5.18a + b** Tree selection from $\hat{\mathbf{s}}$ using the compatibility optimality criterion. **5.15a** The proportion of times that the optimal compatibility tree picked from a random sequence ($\hat{\mathbf{s}}$) of c = 1000 sites had a particular edge in it. The correct internal edge for this tree is edge 7 (marked by the blue diamonds), whereas edges 11 (yellow triangles) and 12 (red squares) are incorrect edges due to parallel and convergent changes. The green circles mark the number of times an unresolved tree was optimal (i.e. there were no observed changes to resolve the short internal edge). Each point represents 40,000 random samples. The weight on each external edge of the model tree is shown along the x-axis. **5.15b** For each sample the sum of the absolute value of entries in $\hat{\mathbf{s}}$ excluding s_0 was recorded. The mean and standard deviation of this sum is shown by the curve marked "all". Likewise the mean and standard deviation of the sum of entries in $\hat{\mathbf{s}}$ not in the tree (s_0 excluded) is also shown (marked "res." for residual). Both standard deviations are small, and each has a scaled binomial variance equal to the expected value of each sum × (1- the sum) / 1,000.

In figure 5.18a, notice the rapid improvement in the recovery of the true tree (7) as rates initially increase. This allows a greater probability of changes on the internal edge, and consequently the proportion of (partially) unresolved trees falls steadily away. The success rate for tree recovery never reaches more than 0.84, a consequence of parallel changes occurring along the long edges, and resulting in patterns supporting s_{11} and s_{12}. As the frequency of these changes increases (especially relative to s_7) the success rate of this estimator begins to fall away. Pattern s_{12} is slightly more probable than s_{11}, which is reflected in the frequency with which each is included in the optimal trees from $\hat{\mathbf{s}}$. We call effects such as these, which are due primarily to differences in the expected value of entries, mean vector effects.

The second part of figure 5.18 shows the sum of absolute values of all entries in $\hat{\mathbf{s}}$ (excluding s_0), and also the sum of absolute values for entries not corresponding to edges in the optimal tree, averaged over the 40,000 replications done at each marked point. We call the second sum the "sum of residual absolute deviations", or just the residual. The sum of observed patterns with the same index as edges in the optimal tree is, of course, equal to the total (all) minus the sum of the residuals. The curved trend for the line marked "all" in figure 5.18b is due to the sum of entries in $\hat{\mathbf{s}}$ being constrained to be one. Meanwhile the sum of the residuals is getting large in proportion to the sum of patterns in the tree; soon it will be as large as the sum of entries in the tree, while for very high rates of change (and thus mutually random sequences) the entries corresponding to edges in the tree will comprise 9/15 or 3/5 of all non-constant entries. Notice also, how small the standard deviations of these sums of deviations are. This is because the sums of deviations are in fact realisations of a binomial random variables, with $c = 1000$. This also explains why the standard deviation of the sum of all absolute deviations begins to drop once the sum reaches 0.5.

5.8.2 Comparative performance of tree selection from $\hat{\gamma}(s)$ and $\hat{\gamma}(d)$

Initially the performance of tree selection on $\hat{\gamma}(s)$ follows a similar trend to that seen with $\hat{\mathbf{s}}$, with the proportion of success rising, while the proportion of unresolved trees rapidly falls (figure 5.19a). However the figure also shows that the maximum success rate does not exceed 0.75, and is always inferior to that of tree selection on $\hat{\mathbf{s}}$, as is shown more clearly in figure later in figure 5.21. We interpret this as being due to the higher variance of the entries 7, 11, and 12 in $\hat{\gamma}(s)$ compared with $\hat{\mathbf{s}}$. This can cause a small but reliable positve entry in $\hat{\mathbf{s}}$ to be transformed to a zero or slightly negative entry in $\hat{\gamma}(s)$ (and thus giving an unresolved tree). Another possibly, also due to the compatibility algorithm scoring ties (something much more likely to occur in the discrete $\hat{\mathbf{s}}$ than the real valued $\hat{\gamma}(s)$) in favour of the true tree containing 7 (Dr Penny is to check for this possibility in his program). Notice that there are more tied unresolved trees being selected from $\hat{\gamma}(s)$. This will be when all three entries take on a negative value, and will also be reducing the absolute success rate for tree selection from $\hat{\gamma}(s)$ (since even taking a guess in such cases would improve the overall success rate). However, even when measuring the success rate conditional on having a resolved tree, selection from $\hat{\gamma}(s)$ does not do as well as selection from $\hat{\mathbf{s}}$ (figure 5.21). The number of times that 11 or 12 was included in an optimal tree, is more nearly

equal than in the case of selecting from $\hat{s}$. This reflects the transformation for multiple substitutions equalising their mean values in $\hat{\gamma}(s)$. As the amount of change increases, the somewhat larger variance of γ_{12} becomes more prominent, seeing γ_{12} selected slightly more often than γ_{11}.

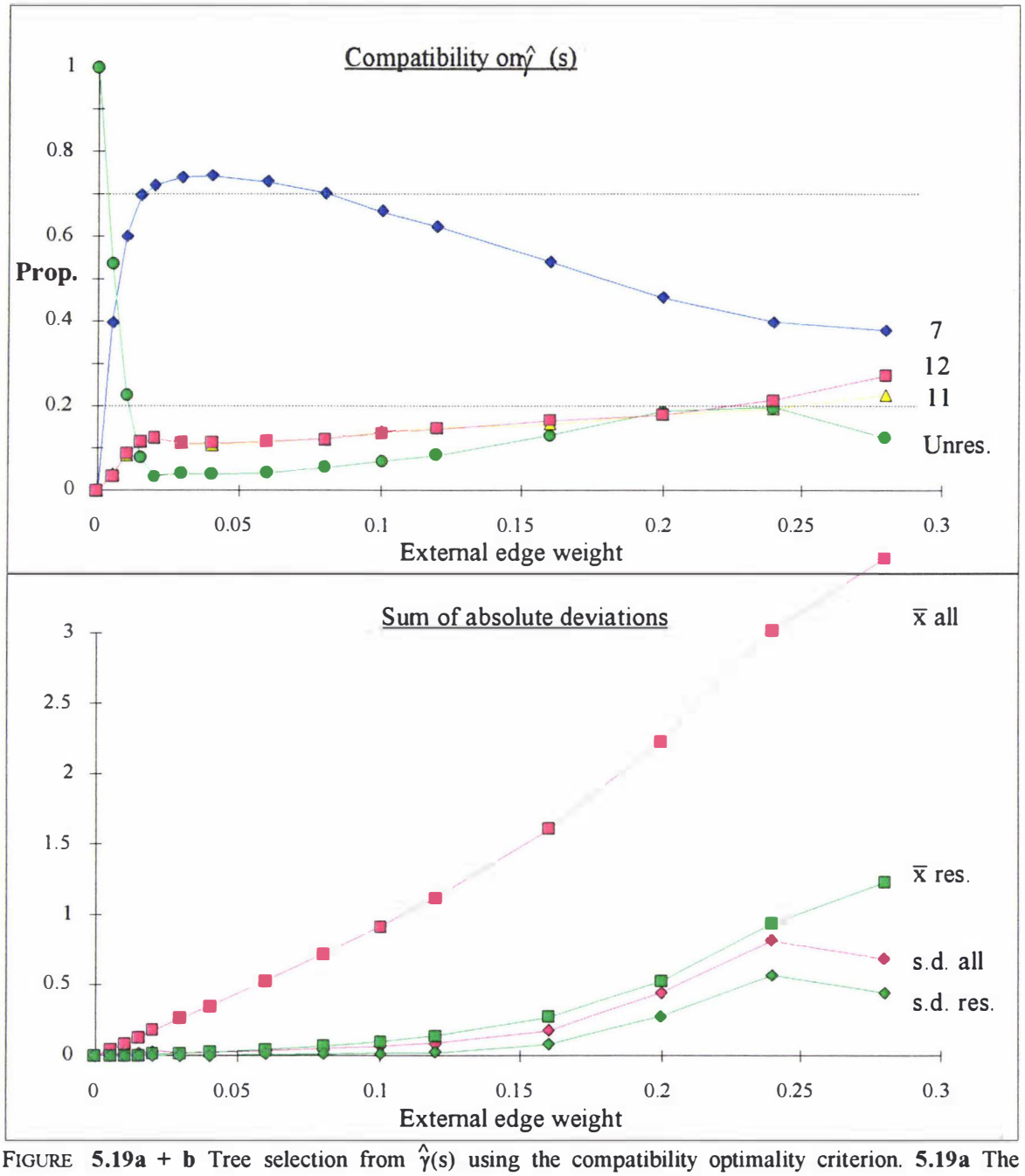


FIGURE **5.19a + b** Tree selection from $\hat{\gamma}(s)$ using the compatibility optimality criterion. **5.19a** The proportion of times that each optimal compatibility tree picked from a random sequence of length 1000 transformed to $\hat{\gamma}(s)$ had a particular edge in it (all labels as for figure 5.18). Each point represents 40,000 samples. **5.15b** The mean and standard deviation of sums of absolute values from $\hat{\gamma}(s)$ ($\hat{\gamma}_0$ excluded) are shown by labeled curves.

As the eight longer edges in the model tree became 0.2 or greater, $\hat{s}$ vectors resulting in occasional negative entries in $\hat{r}$ (hence invalid arguments for the ln transform) became

noticeable. Such samples were discarded and another taken. If they had been included as failures, then the success rate of $\hat{\gamma}(s)$ would have followed more closely the steeper downward trend evident prior to the value 0.2. The drop in the number of unresolved trees at high rates of change was a consequence of not including negative samples in this statistic, combined with the fact that at these large amounts of change the large variances in $\hat{\gamma}(s)$. This allowed for some optimal resolved trees which were more than one partition different to the true tree (i.e. trees that did not have the large internal edges 3 and 15 included in them). Using unbiased estimators (see appendix 4.2) in place of the log transform should improve tree selection reliability in this region, and remove the problem of inapplicable samples.

The sums of absolute deviations from $\hat{\gamma}(s)$ show a quite different trend to those from $\hat{\mathbf{s}}$(figure 5.19b). As expected the sum of the absolute deviations of all weighted entries in $\hat{\gamma}(s)$ initially increases near linearly, with its major contribution being the sum of entries corresponding to edges in the true tree. From approximately 0.12 onwards it begins to curve upwards, due to the greater input from the sum of residual absolute deviations which are loosing their binomial character and getting increasingly large standard deviations due to the transformation. The line showing the standard deviation of the sum (which will be correlated with the sum of the standard deviations of entries in $\hat{\gamma}(s)$) continues upwards until negative values in $\hat{\mathbf{r}}$ cause samples to be rejected. This butting up against a boundary (the rejection zone) then causes the standard deviation to decline, just as it did in the case of single logarithmically transformed pathset lengths (appendix 4.2, table A4.2.2).

The performance of tree selection on $\hat{\gamma}(d)$ (see figure 5.20a) is initially like that on $\hat{\gamma}(s)$, as might be expected by the similarity of their variances for low rates of change (figure 4.15). Plotting the success rate of tree selection from these two vectors side by side (as is done in figure 5.21) does however show $\hat{\gamma}(d)$ to have a higher absolute success rate at the lowest rates of change. One possible explanation is that the slightly larger mean of $\hat{\gamma}(d)_7$ relative to that of $\hat{\gamma}(s)_7$, and perhaps also because the non-tree entries 11 and 12 of $\hat{\gamma}(s)$ are likely to take values larger than 2.0 (see figure 5.15). The slightly more positive means of $\hat{\gamma}(d)_7$, $\hat{\gamma}(d)_{11}$, and $\hat{\gamma}(d)_{12}$ combined with their negative correlations (so that when one goes negative, there is a better than even chance that another has become more positive) results in very few unresolved trees (in fact zero in the simulations for rates higher than 0.01). Tree selection on $\hat{\gamma}(d)$ and $\hat{\gamma}(s)$ does almost exactly equally well at their joint peak (no significant difference at their maxima, as determined by a binomial test). At higher rates of change, the performance of $\hat{\gamma}(d)$ falls away more slowly than that of $\hat{\gamma}(s)$, eventually outperforming tree selection on $\hat{\mathbf{s}}$ when rates of change were high (long edges 0.2 and above). Tree selection on $\hat{\gamma}(d)$ eventually beats tree selection on $\hat{\mathbf{s}}$ because the former vector has mean values closer to the true tree values, and the effect of increased variances is sufficiently low not to fully mask this advantage.

As sequence length gets longer, the variance of both $\hat{\gamma}(d)$ and $\hat{\gamma}(s)$ will decrease until both of them should out perform tree selection on $\hat{\mathbf{s}}$ over all values (assuming that tree selection on $\hat{\mathbf{s}}$ is

not already always getting the correct tree). That tree selection on $\hat{\gamma}(d)$ sees more 11 than 12 chosen at the lowest non-zero rate of change, seems to be due to the distribution of $\hat{\gamma}(d)$ having some distinct skewness towards positive values (figure 4.15). As with $\hat{\gamma}(s)$ at higher rates of change, the increased variance of $\hat{\gamma}(d)_{12}$ relative to $\hat{\gamma}(d)_{11}$ sees it being selected in the optimal tree slightly more often.

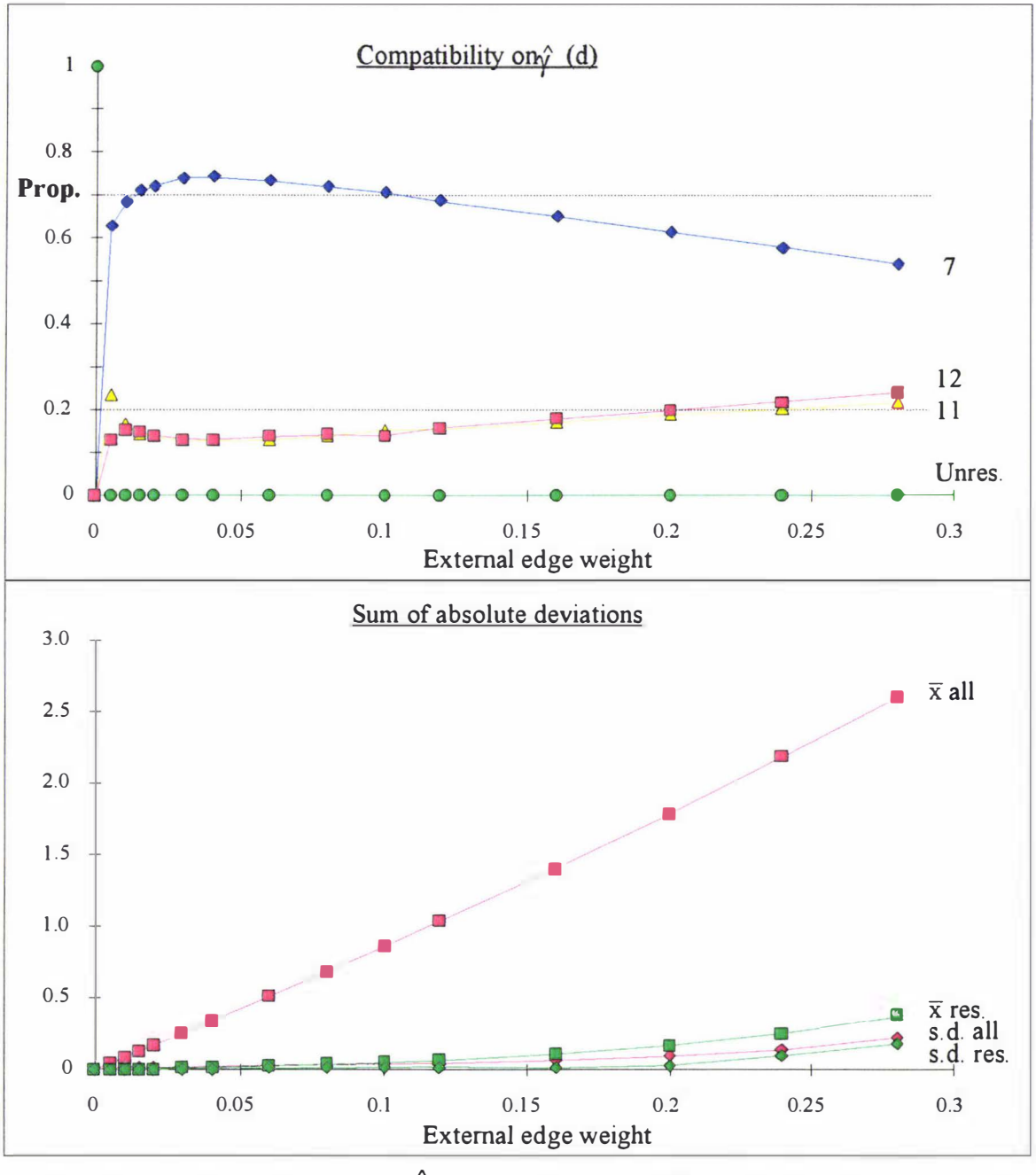


FIGURE **5.20a + b** Tree selection from $\hat{\gamma}(d)$ using the compatibility optimality criterion. **5.20a** The proportion of times that each optimal compatibility tree picked from a random sequence of length 1000 transformed to $\hat{\gamma}(d)$ had a particular edge in it. Each point represents 40,000 samples. **5.20b** The mean and standard deviation of these sums of absolute values from $\hat{\gamma}(s)$ are shown by marked curves (for more detail see caption of figure 5.18).

The trend in the sum of absolute deviations for $\hat{\gamma}(d)$ (fig. 5.20b) is much more linear than that of $\hat{\gamma}(s)$, reflecting the slower rate of increase in the residual term. Indeed the residual of both $\hat{\gamma}(s)$ and $\hat{\gamma}(d)$ becomes lower than that of the residual for $\hat{s}$ for rates above 0.02. The residual of $\hat{\gamma}(s)$ once again becomes larger than that of $\hat{s}$, at about 0.12, but that of $\hat{\gamma}(d)$ remained lowest for all edge length values of greater than 0.02 examined. The residual term of $\hat{\gamma}(d)$ stays about 2/3'ds that of $\hat{\gamma}(s)$, until the latter begins its curve upwards. The standard deviation of the sum of residuals is also small, but often of significant size relative the residual (about 1/3), and the standard deviation continues to increase quite linearly with the residual (figure 5.20b). Here again we see the interplay of the correction for multiple changes in γ acting to reduce the sum of residual absolute deviations (relative to s) (and hence potentially improve the reliability of tree selection). However, the rising errors due to transformation (which are most due to sampling variance) works against this trend, here seen most prominently in the case of $\hat{\gamma}(s)$.

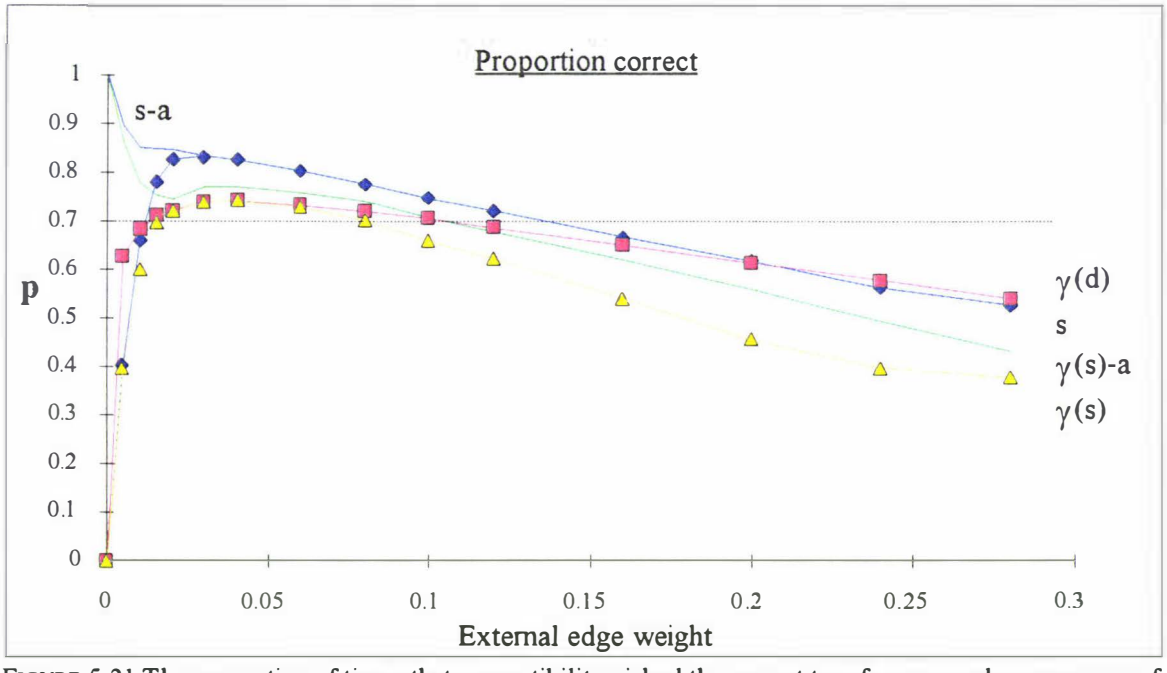


FIGURE 5.21 The proportion of times that compatibility picked the correct tree from a random sequence of length 1,000 transformed as labeled ($\hat{s}$ diamond, $\hat{\gamma}(d)$ square, $\hat{\gamma}(s)$ triangle). The coloured lines without symbols (green for selection from $\hat{\gamma}(s)$ (also marked s), blue from $\hat{s}$ (also marked s-a)), give the probability the true tree was selected given that a resolved tree was selected (for $\hat{\gamma}(d)$ the probability does not change).

It is surprising to see that even for a tree like this, with long edges causing a significant proportion of misleading parallel changes, that tree selection on $\hat{s}$ was most often the best. In addition $\hat{\gamma}(d)$ was seen to generally do better than $\hat{\gamma}(s)$, although we have perhaps biased this by our definition of success, since $\hat{\gamma}(s)$ often did not select a fully resolved tree since all the relevant signals were negative. This is arguably a desirable attribute, expected to occur under the model only when a true edge in the tree is less than 2 standard deviations from zero. It can be argued, we should be most happy with the partially unresolved tree if there is no strong evidence to resolve it (and by its very unresolved nature it should incite researchers to seek a reliable

answer). Due to features introduced to $\hat{\gamma}(d)$ by the minimum of sets of distances criteria for inferring pathset lengths, almost always a tree is chosen due to the positive bias on $\hat{\gamma}(d)_{11}$ and $\hat{\gamma}(d)_{12}$, as well as the lower variances of these entries (see section 4.7.3). Accordingly, when we redefine success as the proportion of times that a method chooses a fully resolved tree and gets it right, then as seen in figure 5.21, $\hat{\gamma}(s)$ out performs $\hat{\gamma}(d)$ up until the external edge weights are 0.1 in length. Tree selection on $\hat{s}$ however still does best by this definition of success. This is because at very low rates of change there is convergence to an infinite sites model, where "two changes are so unlikely that an informative pattern must be reliable". Unfortunately, in reality it is rarely obvious that a site pattern cannot have been due to parallel changes.

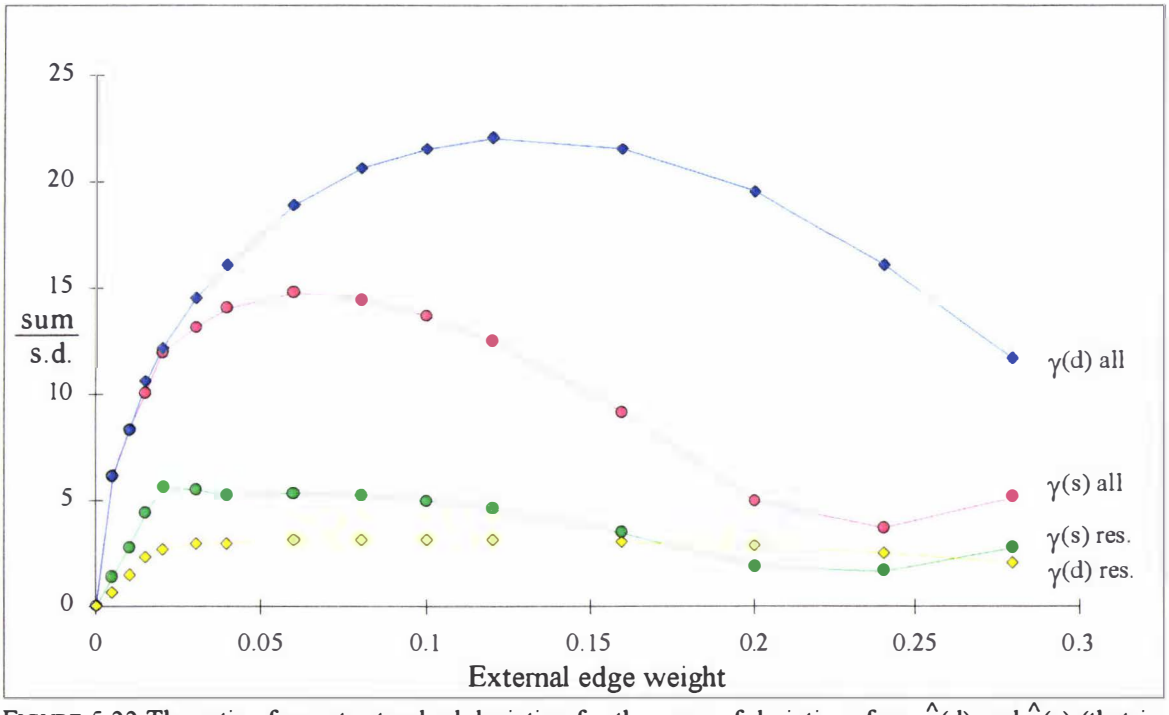


FIGURE 5.22 The ratio of sum to standard deviation for the sums of deviations from $\hat{\gamma}(d)$ and $\hat{\gamma}(s)$ (that is, a signal-to noise ratio). Note that the best signal to noise ratio (of about 22) for the sum of all entries from $\hat{\gamma}(d)$ (which is dominated by the signals corresponding to edges in the true tree) occurs when the external edge weight is approximately 0.13, while the optima for is about 14 and occurs when the external edge length is about 0.065. The average signal-to-noise ratio for edge weights in the distance based trees, is nearly twice that of any single distance through this tree, but for the edges estimated from $\hat{\gamma}(s)$ it is much less (consistent with predictions from section 4.7).

Another feature of spectra, now addressed briefly, is how accurately edge weights are estimated. As already seen, that there is an optima for the recovery of the internal edge that occurs before the actual path lengths reach the point at which their ratio of mean to standard deviation is maximised (e.g. see figure 4.7). Figure 5.22 shows the result of the plotting of the ratio of the sum of all values in $\hat{\gamma}(d)$ to the standard deviation of this sum, to generate a signal-to-noise curve at different rates. Here, this sum is made up almost completely of the edge weights of the true tree; the sum of residuals is small in comparison, and they hardly effect either the mean or the standard deviation of what is mostly the sum of edges in the optimal tree. The peak value

of signal to noise in $\hat{\gamma}(d)$ is at about 0.13, and since the average pairwise distance on this tree is approximately 3 times the length of a external edge, then average path length at this value is approximately 0.39, which translates to an r_i entry of ≈ 0.45. This agrees well with the optima for the curve in figure 4.7.

In contrast to $\hat{\gamma}(d)$, the peak value for $\hat{\gamma}(s)$ (fig. 5.22) lies close to 0.065. At this value it is not the average pathset length, but rather the length of the larger quartets and the sextet which translate to an r_i value of about 0.45. We expect that the difference in these optima for $\hat{\gamma}(d)$ and $\hat{\gamma}(s)$ is due to the longest quartets and the sextet having a much higher sample variance than sums of pairwise distances, as was shown in figure 4.10. The curve of the ratio of size to standard deviation of the residual of $\hat{\gamma}(d)$ appears more flat topped and symmetric than that of $\hat{\gamma}(s)$. This effect may be due to the pre-tree selection of minimum pathsets in $\hat{\gamma}(d)$, which makes the vector $\hat{\mathbf{r}}(d)$ more tree like than that of $\hat{\mathbf{r}}(s)$, so smoothing the residual entries of $\hat{\gamma}(d)$ out. The distribution of sums of residuals for both $\hat{\gamma}(s)$ and $\hat{\gamma}(d)$ is close to normal (not shown), indicating just a slight skew due to the bias of the logarithmic pathset transformation for the largest rates of change.

5.8.3 Tree selection on $\hat{s}$, $\hat{\gamma}(s)$, and $\hat{\gamma}(d)$ when rates at sites are unequal.

It is important to a check the accuracy of tree selection from our the three vectors $\hat{s}$, $\hat{\gamma}(s)$, and $\hat{\gamma}(d)$, when a critical assumption is violated. As earlier, sequences are generated under the 2-state model of section 5.8.1, but with the true distribution of rates across sites following an inverse Gaussian distribution with shape parameter = 1 (c.v. = $1^{-0.5} = 1$). The sequence length used was again 1000, and 10,000 repetitions were performed to estimate each point. Our objective is to see if there are conditions that make the accuracy of the three methods rearrange their order from that in figure 5.21.

Setting the model trees external edge weight to 0.1, we obtain the results $\hat{s}$ (0.68, 0.15, 0.17), $\hat{\gamma}(s)$(0.64, 0.18, 0.18), and $\hat{\gamma}(d)$(0.56, 0.15, 0.29), where the numbers in brackets are the proportion of times 7, 11 and 12 were edges in the selected tree, respectively. This gives a new ordering of the statistical efficiency of tree selection, which is from best to worst, $\hat{s}$ then $\hat{\gamma}(s)$, and finally $\hat{\gamma}(d)$. Notice that two methods, $\hat{s}$ and $\hat{\gamma}(d)$, have become markedly worse performers than under the equivalent i.r. model. The reason that tree selection on $\hat{s}$ has become worse is that the size of $s(T)_{11}$ and $s(T)_{12}$ have both become more similar in size to $s(T)_7$. The decline in the performance of selection on $\hat{\gamma}(d)$, is we hypothesise, due to the way pathsets relating to quartets and sextets are estimated (which involves a type of precursory tree selection, see section 4.7). When distances are underestimated, a short distance plus a long distance will show more underestimation for these mechanisms of substitution than two medium distances. Thus, for our model data with sampling, quartets made up of $\delta_{34} + \delta_{15}$ become relatively more likely than the correct alternative $\delta_{13} + \delta_{45}$ (relative to their expectations under the i.r. model). The size of the quartet {1345} is most underestimated (in expectation), as are other quartets where d34 combined with a long distance (e.g. δ_{16}, δ_{25}, δ_{26}). The reconstructed spectra is then also biased

towards grouping 3 and 4 together (γ_{12}) as the results indicate. The i.r. corrected spectra from sequences, $\hat{\gamma}(s)$, both γ_{11} and γ_{12} take similar values to previously even though site rates are highly unequal. Also helping $\hat{\gamma}(s)$ maintain its performance, are the reduced observed pathset lengths under the i.G. model (relative to the i.r. model), which reduces the sampling variance after the log transformation. So the main result revealed here is the vulnerability of $\hat{\gamma}(d)$ to grouping together a short distance and a long distance during the pathset reconstruction phase, and this resulting in more biased tree selection in $\hat{\gamma}(d)$.

To illustrate the effect of a even more extreme distribution of rates across sites, we keep the same model except for decreasing the Gaussian distribution of rates across sites shape parameter to 0.5 (c.v. = $\sqrt{2}$). Our results are now $\hat{s}$ (0.66, 0.15, 0.19), $\hat{\gamma}(s)$(0.62, 0.19, 0.19), and $\hat{\gamma}(d)$(0.52, 0.15, 0.33). These results are consistent with the trends identified above, that is, slightly more attraction of long edges in $\hat{s}$, more quartet selection problems in $\hat{\gamma}(d)$, and relative stability in $\hat{\gamma}(s)$.

If we now drop the length of the external tree edges to 0.05, but deviate from the earlier model by then making the external edges leading to taxa 3 and 4 0.15 long, the results become $\hat{s}$ (0.31, 0.08, 0.62), $\hat{\gamma}(s)$(0.42, 0.18, 0.40), and $\hat{\gamma}(d)$(0.48, 0.23, 0.29). So, another ranking of the reliability of tree selection from these three vectors arises, specifically $\hat{\gamma}(d)$, $\hat{\gamma}(s)$, then $\hat{s}$ (and while tree selection reliability is poor with the present 1000 sites, the ordering remains stable if we increase sequence length to boost the reliability to over 60%) . Here we see a strong effect of long edges attracting on $\hat{s}$ and a to lesser effect on $\hat{\gamma}(s)$. The crucial bias in $\hat{\gamma}(d)$ is now, relative to the other spectra, more slight than before since the extra long external edges have balanced up path lengths, so that δ_{34} is now nearly equal to δ_{15}, so there is very little bias towards $\delta_{34} + \delta_{15}$ being smaller than the correct $\delta_{13} + \delta_{45}$.

Despite trying many other combinations of tree edge lengths (with the condition that all internal and external edges had the same weight except for the edges 4, 7, 8) and other distributions of rates across sites, no other orderings of the reliability of tree selection from these three vectors was found. When attempting to find a situation in which $\hat{\gamma}(s)$ did the best of the three methods, a logical starting point was the example with all externals 0.1 and the inverse Gaussian distribution with shape parameter 1 (three paragraphs earlier it was seen that tree selection from $\hat{\gamma}(s)$ was doing nearly as well as from $\hat{s}$, under these conditions). However, when increasing the external edge weights to taxa 3 and 4 in order to drop the performance of selection from $\hat{s}$, the performance from $\hat{\gamma}(d)$ increased, converging towards the performance of $\hat{\gamma}(s)$ more rapidly than $\hat{s}$ did. Consequently, $\hat{\gamma}(d)$, rather than $\hat{\gamma}(s)$, became the best transformation to use. Making the overall size of tree larger also favoured $\hat{\gamma}(d)$ due to its reduced variance.

Shrinking all external edges down to a small value (say 0.04), then increasing the edges 3 and 4 to large values (sufficiently large so that they were not just equal to the sum of two of the larger edges, which gives $\hat{\gamma}(d)$ more consistency) looked like a promising area to find the desired ranking. However recovery of the correct tree with the specified sequence length rapidly fell

bellow 0.5, and selection on $\hat{\gamma}(d)$ due to its reduced variance again won out in the instances examined.

In conclusion then, $\hat{\gamma}(s)$ was never the best method in these simulations, and only betters either of the other two methods in somewhat special circumstances. We expect that for biological trees these conditions will be relatively rare (and often undesired input data since the recovery rates are generally very poor). Consequently, as a 'blindfold' method of data analysis $\hat{\gamma}(s)$, seems unlikely be the best starting point for tree selection if the transformations assumptions are approximately met, and sequences are 1000 base pairs or less. Tree selection on $\hat{s}$ was generally most reliable, until it came very close to becoming inconsistent, while $\hat{\gamma}(d)$ tended to do best at high rates of change, especially when $\hat{s}$ was nearing inconsistency. Allowing longer sequences, then in some situations we do expect $\hat{\gamma}(s)$ to outperform the other two methods, as it seems to have a slight edge in robustness over $\hat{\gamma}(d)$, which may only be converted to better tree selection performance as variances drop below present levels. Going to four states should also help reduce variance (plots like that in figure 4.9 versus figure 4.7 show this), since the signal to noise ratio can be up to 50% better, but it will also reduce some advantage due to other methods nearing inconsistency (the reason being that in general, 4-state models result in fewer parallelisms and convergences than 2-state models). This tends to suggest that $\hat{\gamma}(s)$ will find its greatest utility with real studies in visualising trends in the data, and not as a prerequisite for tree selection algorithms. This is not necessarily a bad thing, as there are many methods to select trees, but few methods to analyse signals not in the optimal tree.

5.9 Optimisation and tree selection with URAS

Having developed methods which can accommodate specified distributions of rates across sites (chapters 2, 3, and 5), it begs the question of how we are to use this in a real example. Without any a priori evidence as to the true distribution of rates across sites, then statistical theory suggests optimising some measure of fit between data and model. Earlier in this chapter, are discussed a list of measures between fit of tree and data in order to choose an optimal tree, so these measures are good candidates to trial. Here, are illustrate some relative properties of these different criteria using the 4-taxon rRNA sequences of figure 5.1. These analyses will model a proportion of invariant sites (and a Γ distribution of rates across sites in the case of likelihood only). Invariant sites because they are most simple to understand and to calculate. In addition, this model may turn out to be especially useful when doing maximum likelihood for large data sets as there is evidence to suggest it can approximate continuous distributions of rates across sites with hardly any extra computational burden over the i.r. model.

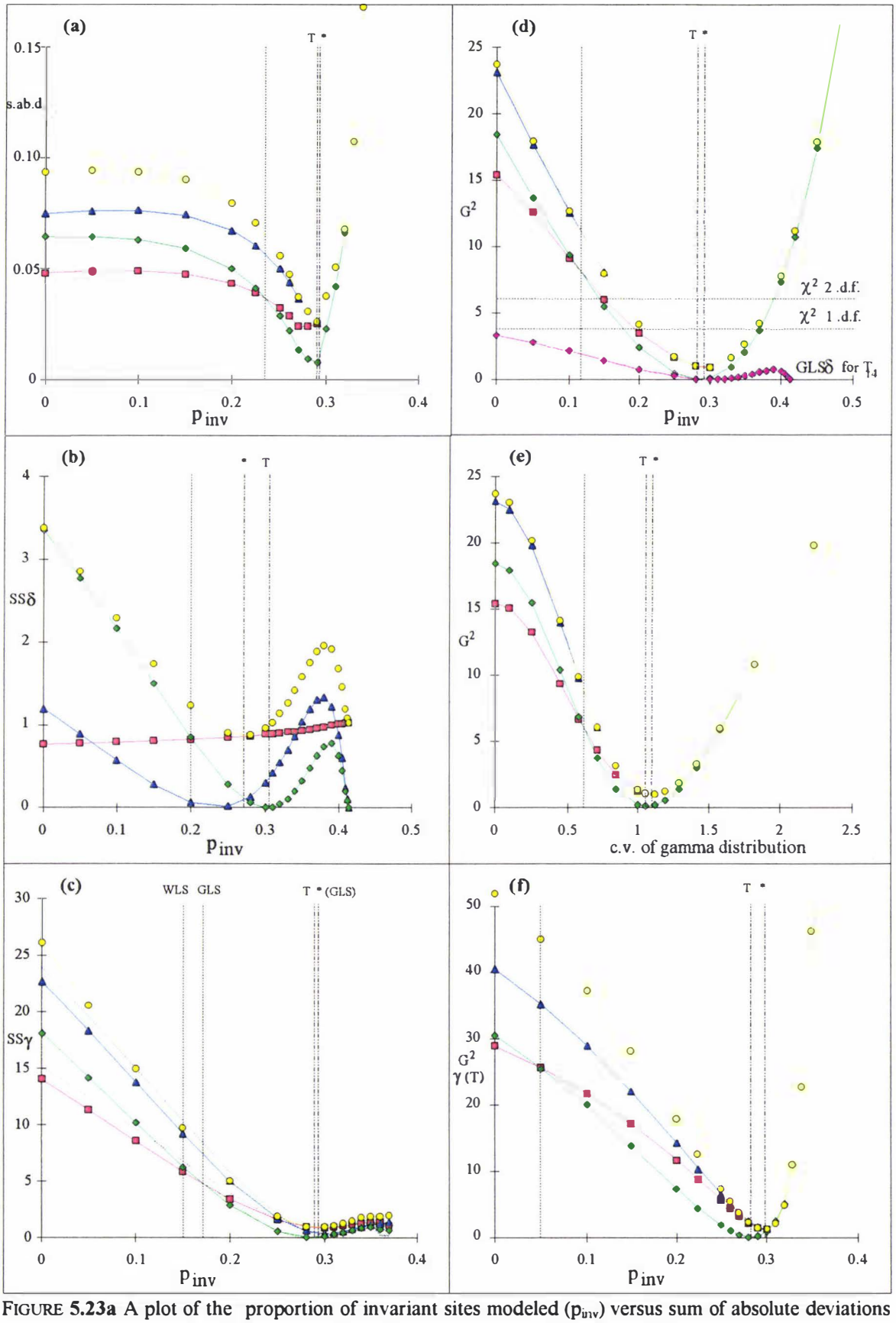


FIGURE **5.23a** A plot of the proportion of invariant sites modeled (p_{inv}) versus sum of absolute deviations (s.ab.d.), for T_{12} (the archaebacteria tree, red squares), T_{13} (halo-bacteria tree, blue triangles), T_{14} (eocyte tree, green diamonds) and T_{star} (yellow circles). The data are the rRNA sequence transversions from figure 5.1. The dotted line shows the point at which T_{14} is more favourable by this criteria than the long edges attract tree, T_{12} The dotted and dashed lines indicate the minima for T_{14}

(labeled with a T) and Tstar (labeled *). **5.23b** As for 5.23a, except that they y-axis is the fit (SS) by the GLSδ method using 2-state Poisson distances. There is a striking double minimum! This is explained in the text. **5.23c** P_{inv} versus GLS SS on $\hat{\gamma}$, shown by lines with symbols, while the associated lines without symbols are the values for WLS$\hat{\gamma}$. The minimum for WLS$\hat{\gamma}$ occurs slightly before that of GLS$\hat{\gamma}$, and like GLSδ, shows a double minimum. **5.23d** P_{inv} versus the iterated maximum likelihood estimate (shown as the log-likelihood ratio statistic, G^2). **5.23e** Coefficient of variation of the Γ distribution ($k^{-0.5}$) versus iterated ML (fit again measured by the G^2 statistic) **5.23f** P_{inv} versus the uniterated log-likelihood fit statistic (G^2), when the tree has edge weights estimated straight from $\hat{\gamma}$ (negative entries in the tree are set to zero).

5.9.1 General trends in fitting a distribution of rates across sites

The results for the optimisation of a "proportion of invariant sites" for various criteria applied to $\hat{\gamma}$(s) is shown in figure 5.23a-d, while figure 5.23e shows comparable results for maximum likelihood when there is a Γ distribution of rates across sites. The fit of all three binary trees (as long as they have a positive internal edge), and of the star tree, are shown. We pay particular attention to three values: the point at which the long edges attracts tree (T_{12}, the archaebacteria tree), becomes less optimal than T_{14} (the eocyte tree) (marked by the dotted line): the point at which T_{14} reaches its minimum: and thirdly, the point at which T_{star} reaches its minimum (these last two points, both marked by dot dash lines). Considering the first value, is

to tie in with our observations on the robustness of each method in the Felsenstein zone, in particular their relative ranking. The value of the distribution of rates across sites at which the minima occurs, is of interest with regards to how it varies between criteria (this data's global minima, across trees, is always on T_{14}). Thirdly, during extensive analyses with maximum likelihood (with the simultaneous maximisation of parameters for a distribution of rates across sites) over the last 3 years (e.g. the analyses for Waddell and Penny 1996), a trend has been noticed. The optimum for the best binary tree always has a less extreme distribution of rates across sites than the optimum with less well supported binary trees, or an unresolved tree (a trend usually most evident with better fitting mechanisms of evolution).

First to consider, are the overall trends, before looking in more detail at each criterion. The values for the proportion of invariant sites at which T_{14} became better fitting than T_{12} (see figure 5.23a-d and f), exactly mimic predictions from section 5.5. The inferred robustness in decreasing order is $G^2\hat{\gamma}$(T), ML, WLS$\hat{\gamma}$, GLS$\hat{\gamma}$, GLSδ, and lastly, absolute deviations or compatibility on $\hat{\gamma}$ (the same as parsimony in this instance). The trend of the star tree to predict a higher coefficient of variation of rates across sites than the optimal tree, is seen for all methods except that of GLS on δ. The amount by which the minima of T_{star} differs from that of the optimal binary tree also follows the order of the robustness of methods, i.e. this difference is least for compatibility on $\hat{\gamma}$ and the most different on $G^2\gamma$(T) (again excepting only GLS on δ). We discuss the probable reason for these trends in the conclusion to this chapter.

Notice also in figure 5.23a-d and f where each curve ends. Those measured on $\hat{\gamma}$ end when just over 0.37 of constant sites are excluded. This is because at this value the similarity measure of the quartet (in the intermediate vector $\hat{\mathbf{r}}$), goes to zero, so it cannot be transformed with the

standard logarithmic method. The method $G^2\hat{\gamma}(T)$ also finishes at a value of just over 0.37, because it can no longer take its tree edge weights from $\hat{\gamma}$. Use of reduced variance estimators would counter this effect. The largest logarithm corrected i.r. distance in (δ), becomes undefined when $p_{inv} \geq 0.41$. In contrast, maximum likelihood does not run into this problem. In fact the iterated minimum G^2 procedure suggests that p_{inv} as large as approximately 0.4, combined with the star tree, still gives an adequately fitting model (according to the chi-square approximation for the distribution of this likelihood ratio statistic under the true model, $\alpha = 0.05$, χ^2 2.d.f. = 5.991, as marked by the horizontal lines on figure 5.23).

5.9.2 Optimising the shape of a distribution of rates across sites using $\hat{\gamma}$ and δ

Here we discuss in more detail the findings with each criterion, starting with fitting between model and the transformed data (either γ or δ). Firstly, figure 5.23a uses the method of minimising the sum of absolute deviations between $\gamma(T)$ and $\hat{\gamma}$, using just the signals pertaining to possible internal edges, in order to optimise a proportion of invariant sites. Close inspection will show that there are in fact two minima by this criterion on all the trees, since the sum of absolute deviations initially (p_{inv} near zero) rises slightly in all cases (this is because the signals for γ_6 and γ_3 initially increase slightly in size, see figure 5.1). This may indicate a potentially difficult problem when using this measure to find the minimum for larger more complicated data sets, where quite different multiple optima might be possible. In addition, the value of the residual sum of absolute deviations increases as the average size of the pathsets increase (as indicated by figure 5.22 above). This effect, if not producing multiple minima, will tend to bias the optimal value of p_{inv} downwards. As mentioned earlier, we don't have a reliable formula for the distribution of this statistic, so without appropriate simulations it is not possible to say for certain at what values of p_{inv} it indicates an adequate fit of the data to the model (although simulations could be run).

The GLS sum of squares on distances, plotted against the proportion of sites treated as invariant is shown in figure 5.23b. The apparent minima of the tree T_{13} (blue triangles), would usually be regarded as 'illegitimate' because this tree has a negative internal edge weight as estimated by this method. The fit of this tree, is included to confirm that its curve is similar in shape to that of the star tree and interestingly it would also appear to underestimate p_{inv}. The archaebacterial tree T_{12} (squares) collapses to the star tree (circles) where their curves first touch, and after that has a negative internal edge weight (notice also that the optimal fit of this tree to the data occurred with zero sites treated as invariant). Only with this GLSδ criterion is the first minima of fit with the star tree coincident with a lower value of p_{inv} than the minima of the optimal binary tree (negative edge weights not allowed).

Perhaps the outstanding result of figure 5.23 is the obvious dual minima with the GLSδ fit criterion. The valid trees, here namely T_{star} and T_{14}, which do not have negative internal edge lengths as p_{inv} approaches 0.37, encounter a second minima just before one of the values in the arguments for the corrected distances becomes negative. This feature also shows up with the

method of WLS on $\hat{\gamma}$, and more subtly with GLS on $\hat{\gamma}$ (figure 5.23c). Since all methods were programmed independently, with different algorithms, this is unlikely to be a computational error; we give our interpretation of its cause next.

The explanation for the double minima occurring in figure 5.23b is that the very large transformed distances, which are encountered just before arguments to the correction formula become negative, also imply very large variances (e.g. see figure 4.7). For all these criteria, with four taxa (GLSδ, WLSγ, GLSγ) every pathset length is involved in estimating every edge length (and every potential edge length, in γ, for any number of taxa). If just one distance is implied to have a variance heading to infinity (so signal-to-noise ratio goes to zero), the variance of all edge length estimates also heads towards infinity. In the case of γ and GLSδ, the SS fit is basically between observed and expected edge lengths (or sums of edge lengths, to give inferred tree distances with δ). This means as soon as one pathset length goes towards infinity, a squared mismatch is divided by increasingly large estimated variances, so the overall sum of squares can decrease. In some cases the variance of all components goes towards infinity, so the overall SS goes to zero, as with WLSγ, (a plot of $\hat{\gamma}_{s.e.}$ versus p_{inv} shows this feature clearly in chapter 6). There are ways to alleviate this problem, which are not guaranteed to remove it (one is altering the relative contribution to overall fit of different components). GLS on γ, for example, by taking into account covariances, does not drop the SS as quickly as WLSγ in this terminal region, although there is a still a clear dip and second minima (initially most apparent on the best fitting tree). WLSδ (not shown) is also affected by this factor, and may show double minima as distances become very large. It is also important to point out, that this feature is not due to the delta method, since this approximation underestimates the variance at very large distances (as long as sequences are sufficiently long that samples are not excluded due to negative logarithm arguments). This feature will occur with all transformations which can assume progressively more unequalness of rates across sites (i.e. the underlying assumed distribution of site rates has an unbounded c.v.).

The occurrence of multiple minima is always a source of concern, in that any optimisation method may converge to either minima. In this case, as long as one of only two minima is on a boundary of the data space (e.g. were distances become infinite) then we may anticipate such a minima, and search for the other minima away from this boundary. For more complicated examples with many pathset lengths, things may get more complicated. Recently Steel (1994b) has shown that multiple maximum likelihood solutions can exist on a single tree. It will be interesting to see if this type of situation can apply also to ML. Figure 5.23, however, shows ML with one clear minima so it may better avoid this problem. Perhaps this is due to estimating fit at the multinomial level where the variances, and covariances contributing to overall fit are quite stable.

Another notable feature of GLS on distances with this data, is that no value of p_{inv} sees the model rejected by its fit statistic (the sum of squares, SS). With enough data the SS has a chi-square distribution under the model (here with 6 - 6 = 0 d.f. for the binary trees and 1.d.f. for the

star tree). Here the data is very saturated in terms of counts per cell, making the approximation to asymptotic normality very good. Consequently this insensitivity is almost certainly a reflection of the distinguishing power of δ, which was also seen to be relatively insensitive in the robustness evaluations made in sections 5.5. Also notice that despite the binary trees having no formal degrees of freedom left for hypothesis testing, these models none the less still contain discriminatory information as one of them (T_{14}) clearly shows the best fit.

The first minimum found by GLS on $\hat{\gamma}$ (figure 5.23c) is in good agreement with the other methods. Its goodness-of-fit statistic, a SS, has reasonable sensitivity to a lack of fit between model and data for p_{inv} less than the first minimum, but this sensitivity then apparently decreases at the highest values of p_{inv}. The explanation appears to be that as path lengths grow large, the SS can become quite small due to the variances becoming very large (and there is indeed a shallow second minima). However, the overall fit in the range of p_{inv} for which $\hat{\gamma}$ is defined the fit is very similar to that of minimum G^2 (maximum likelihood), so it may be that on data without such high rates, these two criteria may behave even more similarly. While WLS on $\hat{\gamma}$ initially shows a slightly larger SS than does GLS, overall the fit by these two criteria is very similar. The SS for WLS on $\hat{\gamma}$ also drops towards zero creating a second minima as p_{inv} approaches the point at which negative pathset lengths occur (this second minima hard to see on this figure 5.23c, but is indicated by the short horizontal blue line).

5.9.3 Optimisation by fit measured at the s level

In contrast to all the previous methods based on fit to $\hat{\gamma}$, optimising p_{inv} simultaneously with the edge weights in the tree by maximum likelihood (minimum G^2) yielded a smooth parabola like curve with only one minima (figure 5.23d). In addition the statistic of fit of data to model which has a chi-squared distribution (with a nominal 1 d.f. for a binary tree, 2 d.f. for T_{star}), asymptotically, under the true model, is generally like GLS and WLS SS on $\hat{\gamma}$, and much more informative than the SS from GLS on δ (also shown on figure 5.23d as the purple line). As the lines for the 1 and 2 d.f. chi-square 95% quantile show, p_{inv} with 95% confidence lies between approximately 0.175 and 0.375 for the best 1.d.f. model (T_{14} plus invariant sites), and between 0.175 and 0.385 for the 1 d.f. star tree plus invariant sites model. These confidence intervals are slightly larger than those calculated by Churchill *et al.* (1992) for the same data (0.201 to 0.384 as p_{inv} under the T_{star} model). This difference is most probably due to their using a slightly different asymptotically justified test. This constructs a 95% confidence interval about the optimal proportion of invariant sites based on the variance from the inverse of the Hessian matrix (section 5.3.2) and assuming a normal distribution (which is approximately the same as constructing a confidence interval on one parameter, without reoptimisation of the other parameters, i.e. edge weights, in the model). This somewhat ignores the contribution of the other variables and tends to make the confidence interval to narrow (see section 5.3.9). The confidence intervals on edge lengths in PHYLIP (Felsenstein 1993, and earlier versions), uses the same test (and has been discussed earlier in section 5.3.7). That this difference is occurring with data which must be far closer to asymptotic conditions than any 4-state data set applicable to the

model in Phylip (having 264 independent data cells to fill) again suggests that these two approaches can give quite different confidence interval widths.

For the ML model of figure 5.23d a confidence interval based on the binomial distribution of the expected number of invariant sites is much narrower, being just 26.0 to 31.0 or just one quarter as wide as previously (so the variance is more than an order of magnitude smaller than that used to construct the previous confidence intervals). This large difference, perhaps indicates how much doubt there is about the inferred proportion of invariant sites due to the inexactness of inferences of edge lengths with this data (the only other free parameters of this model). (While, GLS on $\hat{\gamma}$ may appear less sensitive at locating the optima, it gives a very similar lower limit to ML on the proportion of invariant sites, this being approximately 0.18 by the same asymptotic chi-square distribution test).

Figure 5.23d also shows that all the trees evaluated by ML trees fit the data adequately at their optimal value of p_{inv}. By an asymptotic likelihood ratio test the star tree model would be chosen as offering the best trade of between explanatory power versus fewest parameters in the model (in agreement with the findings of Churchill *et al.* 1992). Application of the minimum X^2 method gave very similar results to that of ML. Here, in agreement with results in the Felsenstein zone (section 5.5) the X^2 statistic seemed to be more sensitive than G^2 in detecting departures to the model for this type of data, giving values that were about 5% to 10% larger than those of G^2 (and it has the same asymptotic chi-square distribution under the true model).

The next figure, 5.23e, shows the performance of maximum likelihood when assuming a Γ distribution of rates across sites. The curve is very like that of the previous invariant sites ML model, and again all trees fit the model adequately. The optimal k value is very close to that inferred in figure 2.7 with a highly similar set of sequences, as aligned by Lake (1988). The optimal fit of this model is fractionally worse than that for the invariant sites ML model (less than 0.2 G^2 units) but the difference is not significant (its barely visible in these figures). In chapter 2 we showed how to constructed models with a mixture of invariant sites, and a distribution of rates across sites for the remaining sites. The ML point for this model, and across all trees, was achieved when k went to infinity (i.e. the i.r. rates), and the proportion of invariant sites took on their optimal values as shown in the figure 5.23d. As mentioned earlier in section 5.3.7, such results were checked by creating data sets sampled under a ($p_{inv} = 0.2$, $k = 1$) bimodal distribution. This seems to be a rather common 'antagonistic' behaviour for this type of mixed rate class model.

The last figure shows results the fit of the $G^2\gamma(T)$ criterion under the invariant sites model. It appears to share the same desirable attributes of the ML method, in terms of having a single optima, and a smooth progression of worsening fits moving away from it. The goodness-of-fit statistic appears to be sensitive, but as mentioned earlier the exact distribution of this statistic under the t e model is presently unknown.

Overall the results show there is grounds for concern over the use of non-ML methods to estimate the number of invariant sites, or equally other distributions of rates across sites (data not

shown). There is a real danger of multiple optima, especially an optima located as the inferred distances begin to get very large, with rapidly increasing variances. The non-iterated likelihood method, $G^2\gamma(T)$, did well, and it will be interesting to see if these methods are generally reliable (e.g. using a weighted least squares distance tree estimation procedure to give edge lengths), as they should be very fast to calculate compared to even ML optimisation on one large tree (e.g. of 50 taxa).

5.10 DISCUSSION

The overwhelming conclusion in writing this overall discussion of results in this chapter, is just how difficult it still is to get an overall understanding of how different tree selection criteria work. It is even harder to be certain of making recommendations of which is best. This chapter goes some way towards answering these questions, but it also makes one aware of the scale of the problem facing the preparation of unbiased recommendations of methods (a problem perhaps ultimately most confounded by our lack of understanding of the some the dynamics of sequence evolution). In considering this discussion, it is important then to separate what we know of the performance of data transformation / tree selection criteria under simple models, from any claims of how they will perform with real sequences. This last question is a field of study in its own right, and requires more concerted efforts to accurately diagnose the processes and features of sequence evolution, and stop assuming they basically fit our models.

An important general finding of this chapter is that all methods can become inconsistent, and they probably tend to do so in unison, making arguments such as "my trees of 3 billion year old molecules are probably correct since I used three simple i.r. i.i.d. methods and got the same result" look rather weak. With this distinction made, the results of this and similar work can be put into a more realistic perspective.

In terms of the tree selection criteria, some general trends did emerge. One of the most important was that the more costly statistical tree selection (e.g. GLS, X^2 and ML) tended to be paying bonuses in robustness (in terms of larger parts of the parameter space yielded consistent tree selection), and the ability to show symptoms that the data was not fitting expectations. In practice, these methods also allow some of the easiest, and perhaps most statistically efficient, frameworks for hypothesis testing. The more statistically intense methods using correlation structure also tended to generally infer the longest edge lengths. Other publications have shown some similar results. The results of Hasegawa and Fujiwara (1993) also show that i.r. ML tends to be more robust than neighbor joining using a simple i.r. transformation, or unweighted parsimony on observed sequences, under non-i.r. / i.i.d. models with 4-taxa. Interestingly, Hasegawa and Fujiwara's (1993) results also suggest the inconsistency of i.r. ML methods when

rates across sites are unequal, although they did not directly discuss this or address it by making exact calculations in place of simulations (they were more interested in robustness). Using simulations, Gaut and Lewis (1995) claim inconsistency of ML based on simulations with short sequence lengths (500 to 2000). Their results are certainly strongly suggestive, but exact calculations were not performed (as verification of inconsistency requires). Our own results showing inconsistency of ML in this region (figures 5.11a-b, 5.13a), were submitted in early 1994, and appear as part of Lockhart *et al.* (1995). One of the useful features of Hadamard conjugations to calculate probabilities, is the very simple structure of the conjugation, and a wide variety of exact test that can be made to confirm it is correct. This is important since there have been a number of cases recently where claims of the performance of ML have been shown to be incorrect due to numerical errors (e.g. see Hasegawa *et al.* 1991).

Extending predictions of robustness to more than 4-taxon trees is very important, since there are some cases where 4-taxon examples give a biased view of what is going on. It will be interesting to evaluate many more methods in the Hendy-Penny zone, especially allowing 4-state data (able to be done with the extended Hadamard conjugations of chapter 2). Our prediction is that the order of robustness will remain essentially the same.

One result which we feel deserves more careful study is just how does GLS on distances compare with GLS on transformed sequences or ML on sequences. Its still surprising just how big a gap appeared between two methods (GLSδ and GLSγ) which differed by only 1 degree of freedom. While our suspicion is that this gap will widen when using more taxa and more states, we wonder if it so applicable in reality. The possible reason, as discussed in the next chapter, is that the effective degrees of freedom of the G^2 or X^2 statistic is vastly reduced by sparseness of the data (i.e. many missing nucleotide patterns). However this sparseness probably affects the distance matrix more slowly, since each distance counts many events at many site patterns. Accordingly, in situations of more taxa, and moderate sequence length, the differences between ML and GLSδ may be minimised.

The findings relating to the anti-Felsenstein zone were particularly interesting. We need to be aware of the fact that models can go wrong in many ways, and unless we can accurately characterise the real process of evolution, it is not possible to be confident of the reliability of current methods. This zone also shows that in evaluating the reliability of a tree selection criterion, it is not sufficient even with only four taxa to consider just the Felsenstein zone (e.g. Huelsenbeck *et al.* 1992), as this can give a very biased view of which tree building methods will be most useful (assuming that the real sequences are behaving in a similar way to the model in the simulations). In such simulations one must use Felsenstein, anti-Felsenstein and narrow internal edge clock like tree to get some balance. Using random branch lengths (or preferably some sort of randomness within realistic expectations) is another possibility (e.g. Charleston *et al.* 1993), but this approach can only reveal an average and not an understanding of important features at work (so is not sufficient by itself). Nor do random edge length simulations give knowledge of which methods are expected to be most reliable in resolving a situation which has two long external edges, while all other are short.

It is even harder to predict what the best method of tree selection might be with six taxa. Our simulations showed that transformation of sequences to correct for multiple changes was frequently deleterious to the chance of success of a method, where all edges were of approximately equal length, and sequences were of moderate length. A variety of conclusions can be drawn from this. One is that we really do need to have good long sequences if we hope that transformations which give mean values over many replications that are reliable.

In contrast, predictive methods like ML and X^2 are capable of both taking account of multiple substitutions, and keeping sampling errors to a similar size to those encountered by other, non-model based methods, which use the character patterns as found (e.g. parsimony or compatibility on the observed sequences). This feature is not only important to the efficiency of tree selection, but also to the reliability of inferring divergence dates. This can be confirmed by running a bootstrap analysis in PHYLIP, using the DNAML 5-parameter model, versus the 5-parameter ML distance and neighbor joining. Upon comparison of the inferred relative divergence times of nodes, the ML method does better, some times markedly so (e.g. in the case of the Horai *et al.* mtDNA sequences, especially when 60% of constant sites are removed from the sequences prior to analysis, results not shown). Recently, Adachi and Hasegawa (1995) have argued that ML methods are less prone than neighbor joining to a downwards systematic error in estimating divergence times. This is consistent with our findings in section 5.2 and 5.3, where ML and minimum X^2 made the largest estimates of the longer edges, given the evidence that the model was underestimating all edge lengths due to failure to account for unequal rates across sites. It was findings such as these, and the lower sampling errors of divergence times that ML methods should find favour in divergence time studies (e.g. Hasegawa *et al.* 1985, Kishino and Hasegawa 1990, Waddell and Penny 1995, Adachi and Hasegawa 1995).

It also seems fair to say that over the evaluations in sections 5.5 to 5.7, and 5.9, that the ML and minimum X^2 methods showed the fewest surprises in their behaviour, and in that sense were reliable. For example, both methods had good robustness in the Felsenstein and Hendy-Penny zones, defaulted to an unresolved tree rather than showing inconsistency in the anti-Felsenstein zone (and did no default until lower values of x), and did not show multiple minima when optimising a proportion of invariant sites. These are especially important properties as the complexity of our models grow.

Its important to keep developing likelihood models because they offer much more flexibility than any other model based method. In this thesis we have looked at the issue that sites can change their intrinsic rate (appendix 2.1, section 5.3.5), and a better understanding of this factor is needed (which is an aspect of covarion type models). Its also important to consider ML models with lineage mixing, something which is certainly important amongst sites in long stretches of nuclear DNA amongst closely related populations (as shown in sections 5.4, and discussions therein). It may yet turn out that reticulate phylogenies are very common amongst "species", or species complexes, of plants, fungi, and basically anything that is not a "typical animal". Such factors will become more pressing as sequences which are more than 10,000 base pairs in length (e.g. with long range PCR), which may contain multiple recombination points. A more detailed

discussion of our assessments of ML models with rates across sites, and recombination, is found sections 5.3 and 5.4.

Development of ML models is probably most compromised at present by a lack of quantitative study of the parameters describing molecular evolution under assumptions other than the i.i.d. neutral model (which really just relies upon average substitution rates) without recombination. Once these are clearly defined it is then possible to consider the best ways to model them. While computational complexity is often cited as a serious hindrance to ML methods, programs such as PAUP* (Swofford 1995) are already showing that analyses with ML models deemed too computationally expensive less than 10 years ago, are now running at a similar speeds (using modern computers) to parsimony in the early 1980's).

Lastly, an important additional application of model based methods is to return reliable estimates of evolutionary parameters. As we saw repeatedly, with present models it is difficult to be confident in factors such as transition to transversion ratio, or divergence time unless the data to model relationship is well understood. An interesting point, is that in traditional model building, statistics like G^2 or X^2 are used to measure fit of model to data, and are a guide to the addition of additional parameters (e.g. see Miller 1990). Under a covarion model, sequences are no longer i.i.d. so in order to measure fit of model to data other statistics may be called for. Another issue is that, generally, it is important to model more than four sequences at a time. For example, the general 3P Γ model applied to a set of four aligned rRNA molecules, returned an apparently reasonable fit. However, this model (like any i.i.d. model) is unlikely to be offering a good description of what has occurred. This is reinforced by tree selection in this instance selecting what is strongly suspected of being an incorrect tree. The lesson here seems to be multiple: Don't believe apparent fit of model to data offers any guarantee of consistency (while its opposite does no mean inconsistency). Avoid measuring fit of models on the smallest possible sets of data, even going from four up to six taxa can make a big difference (ensuring of course that the additional taxa are not joined by relatively short internal edges to taxa already in the tree). Don't let overall i.i.d. likelihood be your only guide of the adequacy of steps in an analysis (from sequencing to alignment to tree selection and selection of parameters in the model). ML is a useful tool, but like all methods there is no guaranteed 'black box' approach.

Given all the uncertainties of Felsenstein, anti-Felsenstein, and Penny and Hendy zones, different trees giving the same sequences, and multiple other uncertainties about the reliability of our models, I believe the approaches used in chapter 3 are sensible in the analysis of ancient molecules. If you have doubt in all current models, you can do much worse than to consider the simple counts of the most conserved sites in light of your hypotheses (e.g. Olsen 1987, chapter 3), rather than throwing caution to the wind, and accepting whatever trees programs give you back.

A5.1 TWO OR MORE TREES CAN PREDICT IDENTICAL SEQUENCE DATA

Under the two state Poisson model with all sites evolving at an equal rate there is a one to one relationship between γ and **s** (Penny and Hendy 1993), and hence each weighted tree predicts different sequence data, which in turn uniquely identify it and its edge lengths. In appendix 2.2 there is a proof that for any specified fixed distribution of relative rates there is also a one to relationship between **s** and γ (by fixed we mean sites don't change their rate relative to each other). However this result does not hold for the whole class of sequences **s** that can be calculated with Hadamard conjugations when we use moment generating functions and their inverses to define the relationship between **r** and ρ. It turns out that the same expected sequence spectrum can be generated by two or more trees given different distributions of rates across sites. It also arises that the same unweighted tree with different edge weights can predict the same sequences when rates across sites vary (in this case the consequence is not an inability to identify the correct tree with certainty, but inability to infer the exact edge weights without precise information on the distribution of rates across sites, and hence problems in estimating relative divergence dates). As we will show later, this second situation can arise when the tree has a number of identical pathset lengths.

A5.1.1 Different trees can give the same sequences: A simple example with 4 taxa.

In table A5.1.1 we illustrate a case in which a 4-taxon star tree gives the same sequences as a 4-taxon binary tree, when the distribution of relative rates across sites varies between the two sequences. The basic form of the two trees used in table A5.1.1 is shown in figure A5.1.1; the unweighted star tree will have edge length ε set to zero, while the unweighted binary tree will give ε a positive weight. The pathset length correction curves for the two different distributions of rates across sites being used here, have already been illustrated in figure 2.6. Because of the flexibility of fitting pairwise distances onto what is effectively a 3-taxon tree, many combinations of edge lengths and distributions of rates across sites will achieve a mapping of the sort illustrated in table A5.1.1. The basic requirement is that the non-linear transformation used with the star tree makes the ratio of δ_{13} to δ_{34} (under in the ρ vectors of table A5.1.1) less than that achieved with another non-linear transform (notice that there are just two different positive pathset lengths on this type of star tree, namely $\delta_{13} = \delta_{23} = \delta_{14} = \delta_{24}$, while $\delta_{34} = \delta_{1234}$, the quartet pathlength). Because of the flexibility of fitting three distances onto a 4-taxon tree, then by altering the proportion of invariant sites, p_{inv}, and the value of k, we can get any ratio of ε to z in our binary tree which is less than 1 (see figure 5.1.1). If ε becomes equal to or larger than z, then we have the impossible situation (for the monotonic non-linear transformations a Hadamard conjugation must use) of $\delta_{13} > \delta_{34}$ on one the binary tree, but $\delta_{13} < \delta_{34}$ on the star tree.

Table A5.1.1 An example where a 4 taxon star tree (described here in vector form) will map to the same sequences as a binary tree

	$\gamma(T_S)$	$\rho(T_S)$	$\mathbf{r}(T_S)$	$\mathbf{s}(T_{12})$ $\mathbf{s}(T_S)$	$\mathbf{r}(T_{12})$	$\rho(T_{12})$	$\gamma(T_{12})$
Index			$r_i =$ $M_{(a:\,0.1,\,k:\,\infty)}(\rho_i)$			$\rho_i=$ $M^{-1}_{(a:\,0,\,k:\,1)}(r_i)$	
0	-1.6	0.0	1.00000	0.48891	1.00000	0.00000	-1.70782
1	0.0	-1.6	0.36152	0.00000	0.36152	-1.76612	0.00000
2	0.0	-1.6	0.36152	0.00000	0.36152	-1.76612	0.00000
3	0.0	0.0	1.00000	0.12739	1.00000	0.00000	0.05830
4	0.8	-3.2	0.23261	0.19185	0.23261	-3.29905	0.82476
5	0.0	-1.6	0.36152	0.00000	0.36152	-1.76612	0.00000
6	0.0	-1.6	0.36152	0.00000	0.36152	-1.76612	0.00000
7	0.8	-3.2	0.23261	0.19185	0.23261	-3.29905	0.82476

Notes to this table: Starting from the left, in the second column, is a vector description of a star tree. Using a Hadamard transform we obtain the vector ρ (third column), which is non-linearly transformed to the vector $\mathbf{r}(T_s)$(fourth column), under a model where 10% of all sites are invariant, and the remainder are i.r. The inverse Hadamard transform, then gives a sequence vector which is not unique to the star tree (in fact infinitely many weighted trees of the star form and a binary form share it). The final column shows one of the countless weighted binary trees (all the same unweighted tree) that share this exact same sequence vector. To show this is so, apply the Hadamard transform to **s** to give $\mathbf{r}(T_{12})$ (of column 6, which is the same as $\mathbf{r}(T_s)$), then apply a nonlinear transform for a model where there are no invariant sites, but the variable sites follow a Γ distribution with shape parameter $k = 1$. This gives $\rho(T_{12})$ (column 7), which are the pathset lengths of exactly one unique weighted binary tree which is obtained in the final column with the inverse Hadamard transform (a diagram of this tree is shown in figure A5.1.1). If we had used a slightly different distribution of rates across sites, a slightly different set of edge weights for this binary tree would be returned.

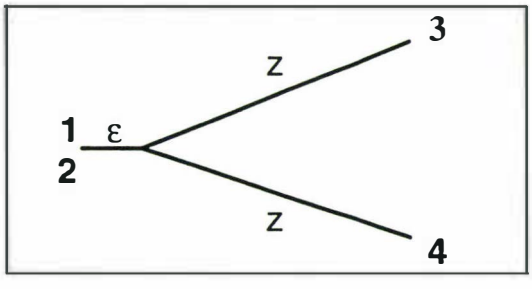


FIGURE A5.1.1 The form of the binary tree in table A5.1.1. Note if we make {2 × edge length z} equal to a point at which our two transformation curves cross (e.g. see figure 2.6), then this binary tree can have the same **s** vector as a star tree (i.e. ε set to zero). To obtain a sequence identical to that of a star tree, start with the same sequence vector, but use a non-linear transformation of the observed sequences which will make δ_{13} equal to exactly 1/2 δ_{34} (whereas on the binary tree $\delta_{34} < 2 \times \delta_{13}$).

A5.1.2 Different binary trees can give the same sequences!

Here, we develop an understanding of the conditions under which distinct binary trees can 'map to' the same sequence pattern probabilities, and illustrate these properties with a series of hypothetical examples.

Our first aim is to understand the conditions necessary for the ρ vector of one tree to map to an **r** vector which can be generated by a distinct binary tree. Because of the orthogonal transformation $\mathbf{H}^{-1}\mathbf{r}$, then if and only if two different topologies can potentially map to the same **r** vector, can they have the same sequence pattern probabilities (i.e. the same **s** vector). A major constraint on the edge weights two different topologies can have in order to get mapping to the same **r** vector is that the non-linear transform converting ρ to **r** must be strictly monotonic under any fixed distribution of relative rates across sites.

Figure A5.1.2 gives an example of one form a tree may take in order to be compatible with an ρ to **r** "dual mapping" given the restriction that the non-linear mapping function must be strictly monotonic. Notice that the branch lengths are all different (in this case a multiple of length *z*), while the internal edge length is an arbitrarily small value ε (much smaller than *z*). When ε shrinks to zero, the binary tree becomes the star tree. Table A5.1.2 shows the tree additive pathset lengths for this star tree (the labeling of tips on the tree is set up to emphasis a clear rank in pathset lengths). From the star tree it is possible to expand out to the three binary trees, i.e. T_{12} = (1,2),(3,4), (and likewise for T_{13} and T_{14}). We will allow each of these three binary trees to have the same external edge weights as those shown on the tree in figure A5.1.2, while the internal edge will be small in all cases (specifically of length ε). Notice that all the trees in table A5.1.2 have the same ordering of pathset lengths as long as ε is less than z. This constancy of pathset magnitude ordering is, of course, essential (for Hadamard conjugation mechanisms at least) in order to map the ρ of one tree onto the **r** of another tree, since the transformations available to do this must be strictly monotonic. Figure A5.1.3 then illustrates how a set of monotonic curves can map these 4 distinct sets of true path lengths onto a single set of **r** values, and hence via the Hadamard transform to a single set of sequence pattern probabilities.

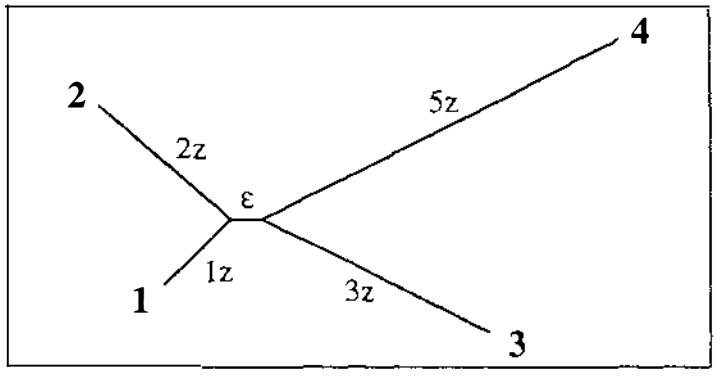


FIGURE A5.1.2. The general from of a weighted binary tree which can potentially map its ρ vector to the same **r** vector (and hence **s** vector) as that of either of the other two unweighted binary trees (which will have similar edge weights) when relative rates across sites vary. (Here *z* is a unit of edge length, so 2z is 2 × z etc., while ε will tend to be a small internal edge, as long as overall rates are low).

Table A5.1.2 The pathset lengths for the different binary trees that share the external edge lengths of the tree in figure A5.1.2. In each case the + denotes that the pathset length on that binary tree equals that of the star tree (column 2) plus a small amount ε (ε < z) which is the length of the internal edge on all the binary trees. Consequently the ranking of pathset length on all these trees is equal. Notice that the pathset lengths are shown in rank of size (and not the order used with the symmetric Hadamard transform).

Star tree		Pathset length on binary tree		
Pathset	length	T_{12}	T_{13}	T_{14}
δ_{12}	3z	same	+	+
δ_{13}	4z	+	same	+
δ_{23}	5z	+	+	same
δ_{14}	6z	+	+	same
δ_{24}	7z	+	same	+
δ_{34}	8z	same	+	+
δ_{1234}	11z	same	same	same
		$= \delta_{12} + d_{34}$	$= \delta_{13} + \delta_{24}$	$= \delta_{14} + \delta_{23}$

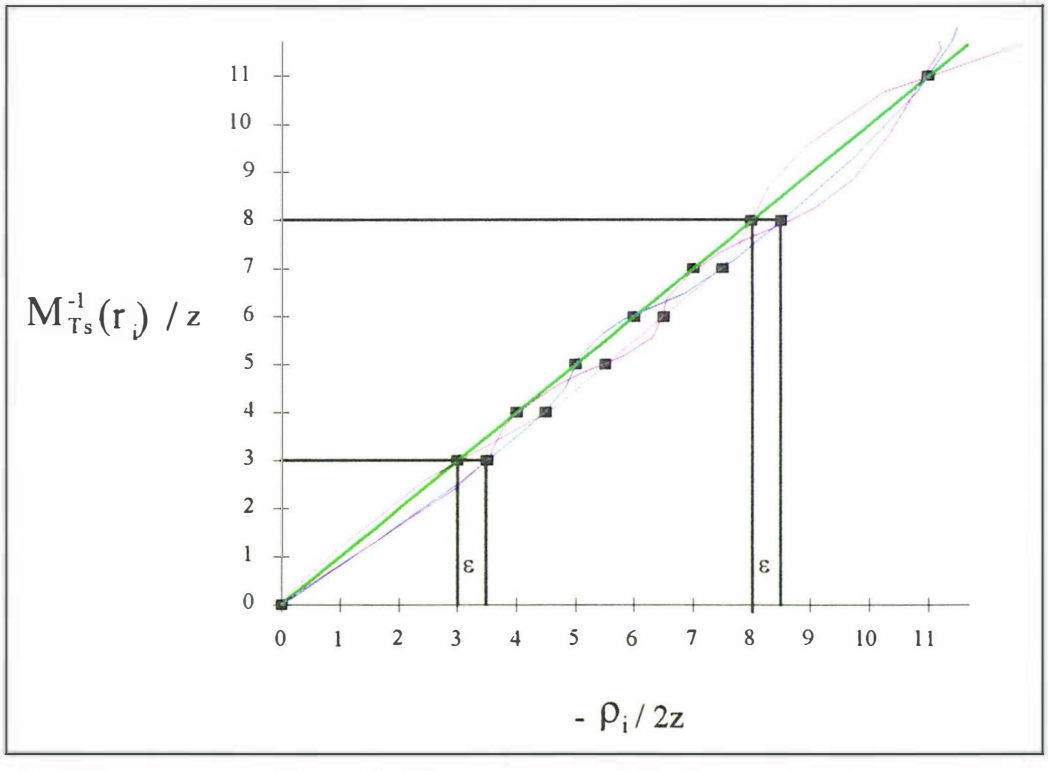


FIGURE A5.1.3 Illustration of the form of non-linear transformation curves necessary to map all four 4-taxon topologies (with edge weights described in figure A5.1.2 and table A5.1.2) onto the same **r** vector, and hence have the same sequence pattern probabilities. The black squares mark all the different pathset lengths on the different trees. The ε represents the contribution of the internal edge to pathset mappings. Each tree has a different coloured transformation curve. The green one is that of the star tree. To tell which tree the red curve belongs to, note that it is implying a longer pathset length (hence +) than the star tree on the 2nd, 3rd, 4th, and 5th ranked pathsets. Referring to table A5.1.2 we see this must be tree T_{12} (given our labeling). Likewise, purple must belongs to T_{13} and the blue belongs to T_{14}. (continued)

Both axes are in arbitrary units of z substitutions per site, the x-axis is linear, while the y-axis has been linearised by use of the inverse transform $M^{-1}(T_x)$(i.e. the inverse of the ρ to $\mathbf{r}$ pathset for the sequences evolved according to T_{star}). (If not linearised this curve would look like a total vs observed distance plot, which is approximately like a plot of $y = \ln(x)$, for $x > 1$)).

A5.1.3 Simplifying 4-taxon binary trees to have fewer unique pathset lengths

There are simplifications of the tree in figure A5.1.2 that are compatible with the monotonicity constraint necessary to allow dual mapping of different topologies onto the same sequence. Looking at table A5.1.2, notice that in all these trees pathsets δ_{23} and δ_{14} are identical, but for the amount z. We can constrain these two pathsets to be equal in length and still get a set of different binary topologies to map to the same $\mathbf{r}$ vector. An example would be to make the external edge leading to taxon 4 (of the tree of figure A5.1.2) equal to 4z giving the new pathset lengths for the star tree of $d_{12} = 3$, $\delta_{13} = 4z$, $\delta_{23} = 5z = \delta_{14}$, $\delta_{24} = 6z$, $\delta_{34} = 7z$ and $\delta_{1234} = 10z$. All three binary topologies, and the star topology, can still map to the same sequence

It is possible to reduce the number of times the pathset length transformations must cross still further, but then only a single binary topologies sequence spectrum may map onto that of the star topology. We now illustrate why this is, with 4-taxon trees forced to have one pair of external edges of equal length and the other pair of external edges equal, while the length of the internal edge (ε) is constrained to be less than any external (figure A5.1.4). Table A5.1.3 shows what the pathset lengths of these trees will be relative to the star tree. Looking at just the first three rows of this table, there are four pathset lengths arranged in an order consistent with a monotonic function mapping ρ_{T12} (the anti-Felsenstein tree of figure A5.1.4) onto $\mathbf{r}_{Tstar}$. Notice that the two monotonic functions achieving this mapping function must cross at least 3 times (since these overall functions are sums of logarithmic functions applied to distinct site rate classes, they cannot be equal except when crossing each other, including at the zero pathlength). This is a reduction from the previous case which implied crossing at least 5 times. Next consider T_{13}. As figure 5.1.4 shows, this tree is like the tree that Felsenstein (1978) used to show that parsimony could be inconsistent (when setting $\varepsilon = x$). Consulting table A5.1.3 it is clear that there can be no mapping of the sequences of this tree onto those of a star tree, since this would require the value in $\mathbf{r}$ corresponding to pathset length $x + y$ to map onto two different ρ values simultaneously! (which is impossible as long as $\varepsilon > 0$). This same impossibility means that tree T_{14} of figure A5.1.4, cannot map its sequences onto those generated by a star topology under any circumstance, and neither can any of these different binary topologies map sequences onto each other (since this would require $\mathbf{r}$ values to be equal but also different in their relative magnitudes, something clearly ruled out by monotonicity).

Table A5.1.3 Pathset lengths for the trees in figure A5.1.4. The + indicates when a pathset length equals that of the star tree (column 2) plus amount ε.

Star tree Pathset	length	Pathset length on binary tree T_{12}	T_{13}	T_{14}
d_{12}	2x	same	+	+
d_{13}	x + y	+	same	+
d_{23}	x + y	+	+	same
d_{14}	x + y	+	+	same
d_{24}	x + y	+	same	+
d_{34}	2y	same	+	+
d_{1234}	2x + 2y	same	same	same

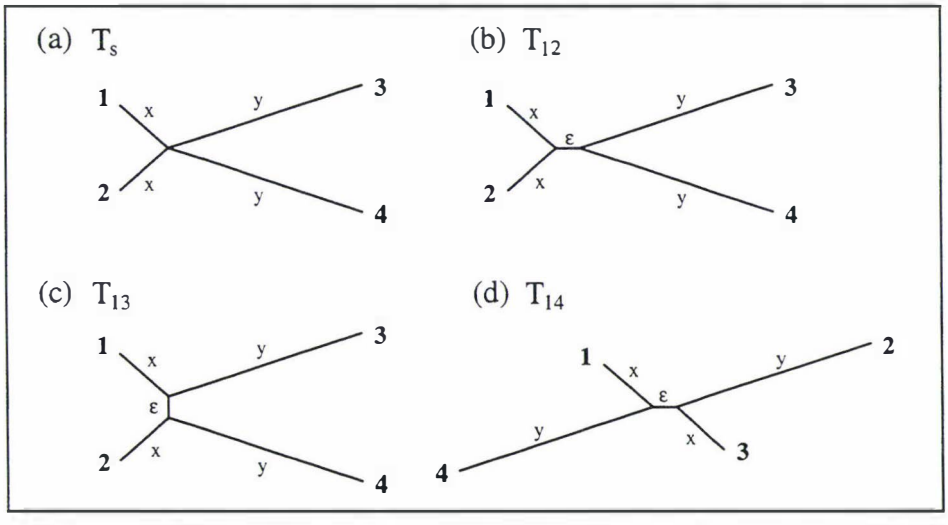


FIGURE A5.1.4. The set of four taxon trees with 2 pairs of external edges forced to have equal lengths. The short external edges have length x, the long external edges have length y, while the internal edge has length ε.

There is a special case of the trees shown in figure A5.1.4 when the length of a pair of external edges is set to zero. This will allow mapping γT_S and γT_{12} to have the same **r** vector, hence sequences pattern probabilities **s**. Figure A5.1.5.1 shows what such a tree would look like, and it is the type of tree we used in our numerical example of table A5.1.1. This type of tree has only three pathset lengths, $\delta_{13} = 0$, $\delta_{13} = \delta_{23} = \delta_{14} = \delta_{24}$, while $\delta_{34} = \delta_{1234}$. As long as $\delta_{13} = \delta_{34} / 2$ (with r_{13} and r_{34} values), it is possible to construct this type of star tree, and then as any transformation such that $2M^{-1}(r_{13}) > M^{-1}(r_{34})$, gives a weighted binary tree with the same sequence vector (where M^{-1} is any inverse moment generating function, see chapter 2). The transformation curves do not even need to cross in this situation. By altering the proportion of invariant sites, p_{inv}, and the value of *k* we are able to get any branch lengths desired on this tree, subject to the constraint that ε must always be less than *z*. If ε becomes ≥ z then we have the impossible situation (when using monotonic transformations) that $\delta_{13} > \delta_{34}$ on the binary tree, but $\delta_{13} < \delta_{34}$ on the star tree.

A5.1.4 What happens with more taxa, or using just pairwise distances?

The condition that the ranking of pathset magnitudes must be the same in any two trees must hold for any number of taxa if different trees are to give the same sequences. If the maximum path length through a tree is fixed (say at 2.0 substitutions per site), and we require that all binary trees map onto the same sequence, then as more taxa are added, the maximum size of an internal edge must decrease. This suggests that the distinction between trees which can map to the same sequence will diminish as more taxa are added. In the limit, as t becomes very large, the necessary correction curves will map to each other very closely, as the pathset lengths will become very tightly spaced, and also because the correction curves cannot be step wise functions (therefore must change gradient more gradually).

External branch lengths might also have to become very unequal in length to guarantee the same ranking of pathsets in the different trees. Indeed t may not need to be very large to make the internal edges need to be very small. Imagine two 10 taxon different binary trees attempting to map to the same sequences, given the constraint that the maximum pathlength through the tree is of a reasonable size, say less than 2.0. Then even under the two state Poisson model, with 2^9 there 512 pathsets, all 45 of which are distances must be less than 2.0 (so too will be many of the other pathsets, if one wants to use unequal external edge lengths to get a clear even ranking of all pathset lengths). To add to this, some of the distances will differ by up to 7 internal edge lengths on the two trees (yet must stay in exactly the same rank order). My guess is that it might be impossible to make the internal edges any longer than, say, 0.0001 each and still get the required mapping of all topologies to the same sequence. This requires further exploration.

If we have a situation where two distinct unweighted trees map to the same sequences, random addition of another taxon (random with respect to assignment of location on a weighted tree and with respect to the length of the new external edge) is expected to negate dual mapping with a very high probability. This is because the points on the correction curves at which multiple mappings can occur have very little tolerance in either direction to get that mapping right (with infinite precision there is no tolerance). This suggests on way of alleviating this problem with real data.

Because distance data has many fewer pathsets (just 45 vs 512 in the previous example with 10 taxa) to be ranked in exactly the right order, then there will be a wider range of parameters under which two distinct topologies can give different sequences, but the same distance matrix. If the distribution of rates across sites is not known, it would then be more likely for a distance method to imply that a perfect fit of model to data had been found, although potentially on the wrong tree! The sequence data, by contrast, would better discriminate that the match of data to expectations was not exact except under the correct distribution of rates across sites. This is a concrete example of how the transformation from sequences to distances can lose information critical to the accurate reconstruction of phylogeny. This is consistent with the expectations of Penny (1982), and answers one of the queries of that paper as to when the difference between just distances and sequences will be important.

A5.1.5 Where can correction curves cross

It is useful to begin looking in detail at the circumstances when the correction curves of two distributions will cross each other twice. With just two correction curves, $f_1(x)$ and $f_2(x)$, the points at which these functions cross each other will be the roots of the function $y = f_1(x) - f_2(x)$ (or visa versa). In this example, (x) may be either the observed pathset length (bounded under this model to have expected value from 0 to 0.5) or corrected pathset lengths (0 to $+\infty$). As explained earlier in chapter 2, when working with **r** entries corresponding to {1-twice the observed pathset length} the range is one to zero, while entries in ρ are equal to {- twice the corrected pathset lengths} and have range zero to $-\infty$. Figure 5.1.1 shows the roots of the function $y = M_1(x) - M_2(x)$, where $M_1(x)$ is the moment generating function (or ρ to **r** transformation) of a distribution with parameters (p_{inv}: 0, k: 1.787625) (i.e. zero invariant sites and the remaining sites following a Γ distribution with shape parameter $k = 1.787625$), while $M_2(x)$ is the moment generating function of a distribution with parameters (p_{inv}: 0.140353, k: ∞). The parameters of these two distributions were chosen because the roots of y are simple decimal numbers -1.2 and -2.4 respectively. We consider that there is also always a root at zero. If M1 and M2 are free to take on any values of p_{inv} and k independently of each other, we hypothesise (from numerical investigations) that:

(1) there are at most two non-zero roots in the region of ρ_i shown in figure A5.1.1.

(2) the maximum x-value a root can occur at, given that there is just one non-zero root, is approximately $\rho_i \approx -1$ (so pathset length would be approximately 0.5 substitutions per site).

(3) when two roots occur with x-values of v and $2v$, the maximum value of v is larger than in case (2), and approximately -1.1.

(4) the minimum x-value of a root in cases where there are two roots is approximately -5, which required extreme parameter values (e.g. $M_1(x)$ (p_{inv}: 0, k: 0.0001) and $M_2(x)$ (p_{inv}:0.99, k:∞)) suggesting there may be an upper bound on the minimum x-value a root can be found at.

As described earlier, in order to map all different binary 4-taxon topologies onto the same sequence vector (and allow each edge to have a unique length) is going to require pathlength correction curves that cross at least 5 times. As just described, it looks unlikely that the invariant sites / Γ model can cross correction curves more than twice (above zero). Thus we expect to have to look to discrete rate classes (probably with five separate rates and different proportions of sites in each class, or 9 parameters each) in order to find monotonic correction curves under our Poisson model that fulfill the role of the coloured lines of figure A5.1.2, and provide a numerical example. A similar numerical optimisation procedure, like that used to find the two equally spaced roots in figure 5.1.5, may be used for this purpose.

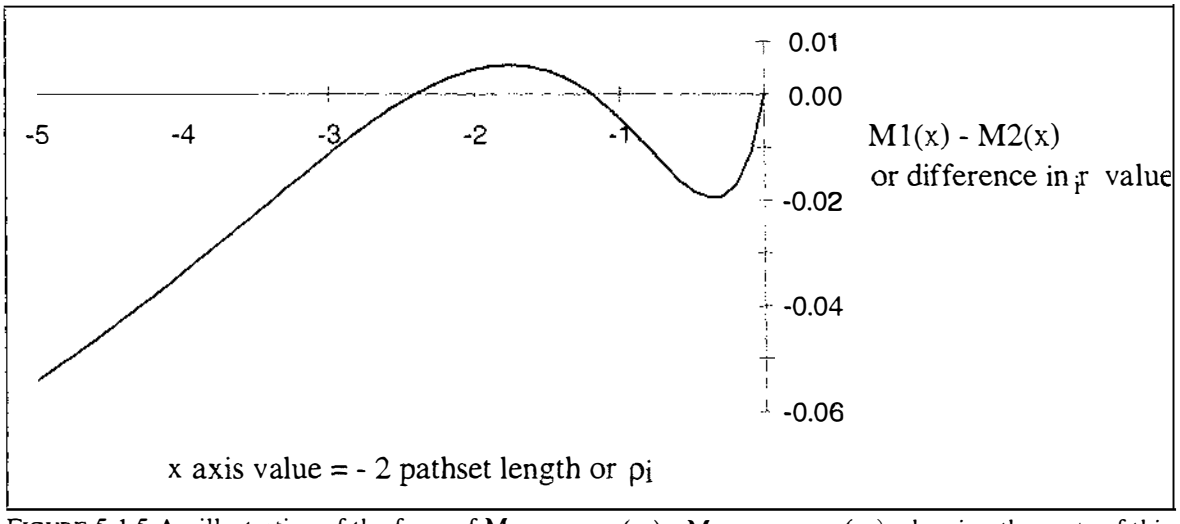


FIGURE 5.1.5 An illustration of the form of $M1_{(a=0.k=1.787625)}(\rho_i) - M2_{(a=0.140353,\ k=\infty)}(\rho_i)$ showing the roots of this equation given that $\rho_i \leq 0$.

It turns out that it is just this sort of discrete i.r. rate class situation that Steel *et al.* (1994) have used to prove an existence theorem that different binary trees (and even all distinct binary trees) can potentially generate the same sequence pattern probabilities. Mike Steel (pers. comm.) was uncertain if the result would generalise to allowing sites in each rate class to have a continuous distribution. Here I theorise it will, and the logic for this is easily shown. If each rate class has a very narrow spread of rates across sites within it (e.g. by a translated Γ distribution with a coefficient of variation e.g. 1/1000, so k = 1,000,000), then this will make very little difference to the correction curves shape, as long as e is moderately large, the correction curves can still cross as required. However, as the distribution of rates across sites in each class spreads out, we expect that the correction curves for each different mean rate class will become more similar, thus forcing the maximum size of the internal edges in the binary trees to be mapped to the same **r** vector to decrease.

A5.1.6 Discussion

These examples are interesting in that they show that the sort of inconsistency problems encountered with long edges attract etc. have an even more interesting conclusion of not just inconsistency, but potential indeterminacy. It is the inconsistency problem that appears as though it will cover a much wider part of the parameter space for any given model, while the different trees mapping onto exactly the same sequences will be special cases.

It would be unwise to dismiss these examples as irrelevant, as they have short internal edge lengths, and require precision to occur exactly. After all this is how Felsenstein's (1978) work was often initially regarded, because the parameter values seemed unrealistic. It will be interesting to see who much of a problem to models having not just identical, but also nearly identical sequences, will be under more general i.i.d. models where more things can fluctuate (e.g. base composition, the transition rates on segment of an edge). Another factor which could elevate this problem to allow larger internal edge lengths in the problem trees, would be if non-i.i.d. models (e.g. some sort of covarion model) imply non-monotonic functions between observed sequence mismatches, and tree additive distances. Also, when working with finite

sequences, nearly identical pattern probabilities may make it practically impossible to differentiate trees given finite sequence lengths, unless we have some other source of information on the distribution of rates across sites. This interesting new area of study should caution use about the confidence we assign to the trees we obtain, and realise there is much we have yet to understand about not only the real process of evolution, but even the properties of the simple models we deal with.

An interesting implication of this work, is that it clearly shows that adding extra sequence does add important information: it eliminates dual mapping in most cases, or alternatively cuts down the number of distributions of rates across sites which could give dual mappings, and can also force internal edges to be smaller for dual mapping to still occur. Extra sequences are effectively giving us more information about the rate of site i relative to site j, and only due to uncertainties in this information, can dual mapping occur in these examples.

Another problem which arises is that different weighted trees (but the same topology, or unweighted tree) can give the same sequences. In such cases, without knowing the true distribution of rates across sites it will be impossible to pin down important parameters in the model. These could include tr / tv ratio, edge lengths, and hence also relative divergence times. We wonder if the similar fit of otherwise matched invariant sites and Γ models on the mtDNA data of section 5.3 might be close to this type of situation.

We end this section with a short series of questions and conjectures:

Question: How many free parameters must the distributions underlying our correction curves have in order to map a specific type of tree onto another sequence spectrum?

Conjecture: The number of times the correction curves can potentially cross does not exceed the number of free parameters in the distributions underlying these correction curves.

Conjecture: The total number of different edge lengths in the two trees being mapped onto the same sequence pattern probabilities must always be less than, or equal to, the number of free parameters in the two correction curves.

Conjecture: If we make all edges in a binary tree take different weights, then for such a tree to map onto every other similarly specified binary tree will require that the correction curves have at least 2^{t-1} free parameters (under the 2-state Poisson model).

CHAPTER 6:

STATISTICAL TESTS

6.1 INTRODUCTION

It is important to be able to make decisions even when there is uncertainty. Statistical tests offer an objective and powerful way to make these decisions. This section presents statistical tests to answer the sorts of phylogenetic questions we are most interested in. Many of these tests are new applications of statistical theory to phylogenetics, while others draw on approximations specific to phylogenetics. We do not review all the tests that have been proposed by others, or used in phylogenetics; for this purpose there are recent reviews, especially Felsenstein (1988, 1993), Li and Zharkikh (1995), and Swofford *et al.* (1995). This section is concluded with a section on Bayesian statistics, and their potential usefulness in phylogenetics.

6.2 OVERALL FIT OF DATA TO THE MODEL

Before going on to make an extensive analysis, its important to feel that the analysis will have some meaning. One way of doing this is to be confident that the assumptions that the method is making are, by and large, correct. There are many ways of doing this. The most general way is to understand something of how the real data, in our case the process of evolution, could mislead methods of analysis. One of the first models put forward in molecular phylogenetics was by Jukes and Cantor (1969) who wished to have a method that was more robust to parallelisms and convergences than using Hamming (observed) distances. They did this with a simple logarithmic correction, which is now known to be an ML estimator of an additive distance (the average number of substitutions per site) under a simple Poisson model e.g. see Zharkikh (1994). This same process continues today with the incorporation of more biological knowledge into our models. A good example is Dr Peter Lockhart's overwhelming concern for unequal base composition causing artifactual grouping of sequences with all methods (see Lockhart 1990, Lockhart *et al.* 1992a, 1992b), leading to the development and application of the LogDet method (Lockhart *et al.* 1994, Lake 1994, and see also section 3.4.4). In the section next section, statistics which can help us to measure violations of the model are considered. In themselves they are not a substitute for studying the actual process of evolution, but can serve as a gauge for refinement of a method.

6.2.1 Measuring overall fit

In traditional i.i.d. models one summary way of evaluating 'adequacy' is by a general fit statistic between the observed and predicted data. Since the data is assumed to be a random sample from a multinomial distribution, then with a large amount of data sampling errors become multivariate normally distributed (asymptotically as c the number of sites $\rightarrow \infty$) . In turn, if data comes from a multivariate model, then the sum of squares of errors between observed and

expected data will distributed as a chi-square (χ^2) variable, with a 'spread' (called the degrees of freedom, or d.f.) that will be dependent upon how many independent variables there are within the model, and how parameters of the model are estimated.

Under the multinomial model with x categories (or cells) and a fixed sequence length, there are x - 1 degrees of freedom available before any parameters are estimated (a degree of freedom is 'lost' since the observations must sum to c). Hence, d.f. = x -1 also as the multinomial converges to multivariate normality (Stuart and Ord 1990, p. 1159). Every parameter efficiently estimated from the data, drops the degrees of freedom. Asymptotically, the observed decrease in degrees of freedom per efficiently estimated parameter is in many instances exactly 1. If the estimator is not statistically efficient, then there need not be a simple bound on what the overall degrees of freedom are. If an estimator tends to perform much worse than an efficient estimator, then there overall degrees of freedom of the fit statistic (SS) can increase over that expected with no parameters fitted. The distribution of the fit of **s**(T) to the observed data, when edge weights in **s**(T) are taken straight from a Hadamard conjugation probably fits the scenario where the degrees of freedom of the fit statistic may exceed x (another example is given by methods of moments generators, Stuart and Ord 1990, 1171).

If the data does not come from the proposed model, then asymptotically its sampling and systematic errors when forced to an incorrect model will have a non-central chi-square distribution (e.g. Stuart and Ord 1990, p. 865). The important thing about this distribution is that the sum of squares of errors is expected to be larger than under the true model, and this is because squared errors will not only contain the random component but also a systematic error component (the size of this 'displacement' overall defines the non-centrality parameter of the non-central chi-square). Consequently, one way of hopefully moving towards a better model (a closer approximation of the true model) is to keep looking for models with lower sum of squared errors. It would also make sense to only allow in parameters which drop the fit by more than 1, since it is expected that efficient estimation of a parameter in the model will achieve at least this decrease.

Under the multinomial model there are many ways of measuring the closeness of fit between data and model. One whole family of statistics are known as the power divergence statistics (Read and Cressie 1988). If we measure the fit between data and model as,

$$\text{Fit} = \frac{2}{\upsilon(\upsilon+1)} \sum_{i=0}^{x} observed_i \left[\left(\frac{observed_i}{expected_i} \right)^{\upsilon} - 1 \right] \qquad (6.2.1\text{-}1)$$

we have the power divergence family of statistics (where υ is a real valued parameter chosen by the user). In equation 6.2.1-1, cases of $\upsilon = 0$, and $\upsilon = -1$ are defined as the limits $\upsilon \rightarrow 0$ and $\upsilon \rightarrow -1$, respectively. One member of this power divergence family is the G^2 statistic (the G statistic of Sokal and Rohlf 1981), which as already seen in section 5.3.1 is a likelihood ratio statistic. Another close relative is Pearson's X^2 statistic. Setting υ to 0 (the limit as $\upsilon \rightarrow 0$) in equation 6.2.1-1, we obtain the G^2 statistic(Σ observed ln(observed/expected)), while setting υ to 1 gives

the X^2 statistic ((observed-expected)^2/expected) summed over all cells. Interestingly, asymptotically under the model, all members of this family are BAN (best asymptotically normal estimators) which makes them statistically efficient (Read and Cressie 1988). The likelihood statistic has the property of being most powerful (capable) of detecting global (evenly distributed) systematic departures from the model, while the X^2 statistic is better at detecting local departures (some cells deviating more than others).

Many people think of likelihood as being the most efficient statistical estimator with no qualifications. This is untrue. Maximum likelihood is generally only proven to be the most efficient asymptotically (where under the i.i.d. model it is not alone in this property). For finite amounts of data (finite sequence length) maximum likelihood often performs well, but is not necessarily optimal (i.e. has minimum sampling variance)(we saw a clear demonstration of this in appendix 4.2 with regard to distance estimation, see also Kuhner and Felsenstein 1994). Our studies also suggest that in some particular instances (where rates across sites varied, but the model assumed they didn't) the X^2 statistic was more powerful at detecting departures from expectation, which gave rise to more robust tree selection than likelihood in the Felsenstein zone (see section 5.5, a trend with was slightly reversed in the anti-Felsenstein zone of section 5.6). Overall, these two statistics are expected to perform the best of the power divergence family (although the intermediate between the G^2 and X^2 statistics obtained by setting $\upsilon = 2/3$ also has some refined properties, see Read and Cressie 1988). The likelihood statistic has a clear advantage over all the other power divergence statistics when analysing sequences in that it is not necessary to calculate all pattern probabilities to infer its value (since unobserved patterns contribute zero to this statistic, see section 5.4). This is would appear to be very important in most analyses of 4-state data with more than approximately 12 taxa.

6.2.2 Factors which distort the asymptotic distribution of fit statistics

Here we identify four factors which will distort a goodness-of-fit statistic, so that its true distribution under random resampling of the data can be quite different from the usually assumed asymptotic form (i.e. χ^2 distributed with d.f. = x - p - 1). In decreasing order of possible severity for these factors are:

(1) Sparseness of nucleotide sequence data

(2) Iterative selection of model parameters

(3) Tree selection

(4) The type of estimator used

The first factor is a well known and has been studied in more classical situations. Sparseness means that expected cell counts will tend to be small (less than 5 and more severely less than 2), and a good discussion and summary of results on this is given in Read and Cressie (1988). If expected cell counts are small, but even (say expected values of about 2), then both the X^2 and G^2 statistics still have a distribution close to their asymptotic form. However, this finding does not necessarily hold for sequence data where expected cell counts are also highly uneven (e.g.

0.8 of all sites might be 'constant', but only 0.000001 of all sites might be expected to have the pattern AGCTAG). The potential for this problem has long been recognised when using likelihood ratio tests in phylogenetics (e.g. Ritland and Clegg 1987). Recently Reeves (1992) and Goldman (1993a) have shown that the problem can be severe with as few as four taxa. They did this by inferring the expected G^2 distribution with Monte Carlo simulations. In many cases, the G^2 for sequences of the lengths commonly used in phylogenetics had an expected value that was only a fraction of what it would be under multivariate normality. What this means is that a fit of data to model with a G^2 of say only 1/3 the expected value asymptotically (and on this basis an incredibly good fit) can actually turn out to be an incredibly bad fit. A small part of this effect might be due to ordinary ML estimators recovering some 'lost' degrees of freedom (Stuart and Ord 1990, and discused below under factor (4)).

Following the 1993 studies in Waddell and Penny (1995) we checked overall fit of data to model in a less extensive way than a full Monte Carlo simulation. This 'sampling' does not involve searching for a tree, but rather considers just the distribution of resampled data under the model without fitting any parameters. This was done by drawing random samples from the expected pattern probabilities under the best fitting invariant sites model. As figure 6.1 shows, even this distribution (which neglects the improved fit expected by the fitting of the 22 parameters of the 5-taxon generalised 3PΓ or p_{inv} model) clearly rejects the data coming from the model. Here, sparseness is causing the problem, and we suspect that usually it will be the major cause of the G^2 statistic being nothing like its asymptotic form.

An interesting point, is that in figure 6.1 the number of unique patterns in the data (94 of the possible $4^{t-1} = 4^4 = 256$) falls very close to the mean of the resampled data. We wonder if this might turn out to be a repetitive pattern, and if so may serve as a rough guide to how well fitting a model must be to approach the expected fit given sparseness. This point has some merit, since patterns which are not seen in the data can add nothing to the G^2 statistic (and typically only a small amount to X^2). Further, McCullagh and Nelder (1989) show that under extreme sparseness the expected mean fit of the G^2 statistic to the data is none other than the number of observations (which becomes equal to the number of distinct patterns since no cell is expected to have more than one count). Being so easy to evaluate for on any sized tree, this possibility deserves further study.

Iterative selection of parameters (factor 2) relating to the mechanism of evolution has not previously been recognised in phylogenetics as a problem in distorting the goodness-of-fit statistic, but is has the potential to do so (and possibly more severely than tree selection). This is a well known problem in other areas of statistical modeling (e.g. Miller 1990). It is easy to envisage that as more extensive programs become available in phylogenetics, many users will step through many possible combinations of model until they find a best fitting one. Furthermore, the number of possible models could become very large (e.g. even the generalised Kimura 3P model has over 5^{2t-3} restrictions possible by reducing the number of parameters per edge only on certain edges). This sort of distortion is very hard to counter or evaluate, and will not be picked up in the Monte Carlo simulations used by Reeves (1992) and Goldman (1993a). To detect it one

would need a more extensive Monte Carlo. This would involve sampling data from a best fitting model found, then implementing a whole cycle of tree selection, model evaluation and stepwise model improvement. Repeating this many times over would give an estimate of how the fit statistic is really distributed, and perhaps most critically, how much the average goodness-of-fit statistic was reduced compared to tree selection fixed on the best fitting model.

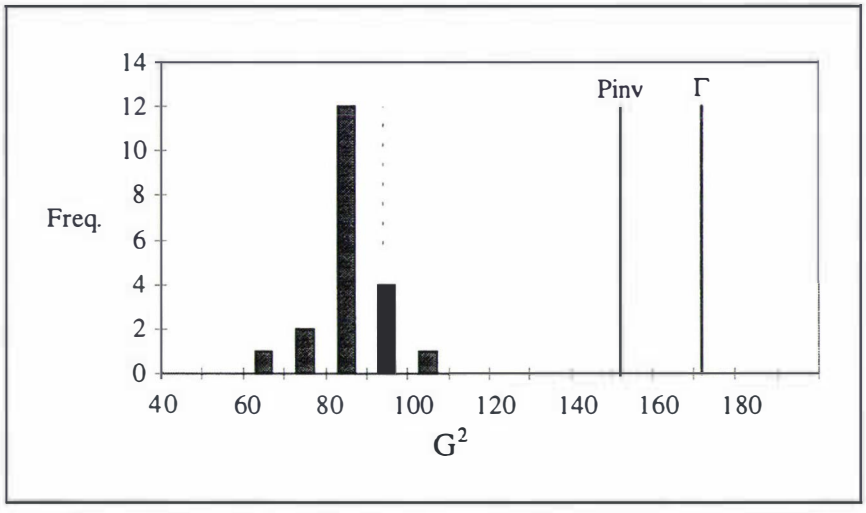


FIGURE 6.1 The distribution of the G^2 statistic of resampled data (no parameters optimised) from the expected pattern frequencies of the best fitting generalised Kimura 3ST model with p_{inv} to 5kb mtDNA sequences of human, chimp, gorilla, orangutan and siamang (Horai *et al.* 1992, section 1.9.2). Notice that the best tree models (the 3P p_{inv} and Γ model) do not come close to fitting as well as this expected distribution (which must fit worse than the statistically efficiently optimised tree models). The vertical dotted line indicates the number (here 94) of distinct Kimura 3P patterns shown in this data. The asymptotic approximation would suggest that the G^2 statistic will be distributed χ^2 with 256 - 1 - 22 = 233 degrees of freedom. In these analyses, the X^2 statistic usually had a value of approximately twice that of the G^2 statistic on the better fitting models (although the relative fit of G^2 to X^2 fluctuated wildly amongst Kimura 3ST submodels). Since the data came from a tree with specific parameters, the sampled G^2 statistic is the same as the goodness-of-fit on that tree without reoptimisation of edge weights, distribution of rates across sites, etc.

Both Reeves (1992) and Goldman (1993a) noted that tree selection would be having a part in reducing the goodness-of-fit statistic beyond evaluation of the model on the true tree (if known). The degree by which this occurs will depend a lot upon how strongly the data 'suggests' a single optimal tree. If we take the Horai *et al.* (1992) mtDNA data, then the tree practically picks itself under almost any tree selection procedure (i.e. bootstrap support on all internal edges is very high, in this case over 99%). Thus there is little chance of getting an alternative tree. With most data, however, there will be a good chance that the best fitting tree in a random sample is not the true tree. As Reeves (1992) and Goldman (1993a) note, their Monte Carlo simulations will take this factor into account, but they did not evaluate its size relative to the sparseness problem. We suggest its effect will be relatively small since tree selection has a strong hierarchical structure, and when there is the opportunity to pick an alternative tree, it will usually be around internal edges which make little difference to the overall likelihood. If there is a large chance fluctuation in a few cells in the data, it seems unlikely even one of these will be compatible with a poorly resolved edge in the tree, so tree selection is not free to roam over the data to eliminate these

fluctuations from contributing lack of fit to data. However, fine grained stepwise model optimisation could possibly do this more effectively (hence the suggestion this may be a more serious problem at least with few taxa). Network selection (a generalisation of tree selection) though, could be a serious problem for the reticulate phylogeny methods of von Haeseler and Churchill (1993). These are not constrained so rigorously by hierarchical structure as the reticulate phylogenies presented in section 5.4.

However, hierarchical structure will not always so effectively reduce the effect of tree selection on the goodness-of-fit statistic. Under large, poorly resolved phylogenies (e.g. intraspecific sequences), it could become a major factor. In these sorts of situation, there may well be much more choice (and hence decrease in the expected goodness-of-fit) amongst the edges included in the optimal tree, versus the parameters chosen in the mechanism of evolution (which may well be few and simple for reasons of computational burden).

If tree selection is not such a big issue, it is reasonable to speed up maximum likelihood Monte Carlo simulations by fixing on the unweighted optimal tree (but not edge weights)(a suggestion made to Nick Goldman when he sent a preprint of his 1993a paper). Another alternative would be to identify a small set of trees most frequently occurring in resampling (perhaps all trees within a few G^2 units of each other on the real data) and evaluate just these trees, rather than the larger investment of an extensive search strategy. These issues require further study. However, since these statistics are just a guide, then concern for fractional improvements of the power of our tests may not be of much importance (e.g. we know the models are 'wrong', so does it really matter if in one sample model *a* does fractionally better than model *b*?).

The last factor (4) is theoretically expected to be worst with few cells, but a relatively high proportion of parameters estimated. It turns out that with ML estimators (and other efficient estimators) the asymptotic goodness-of-fit need not be χ^2 with degrees of freedom $x - p - 1$ (where x is the number of cells, and p is the number of parameters optimised). This is due to a factor known as partial recovery of degrees of freedom with ordinary ML estimators, so that the asymptotic χ^2 distribution can have degrees of freedom between $x - p - 1$ and $x - 1$. It is possible to estimate this factor (Stuart and Ord 1990, p. 1169), although in practice this can be time consuming, computationally expensive, and not always reliable. The usual approach is to resort to making a test assuming either possible extreme of the distribution and hoping for agreement in both cases. Partial recovery of degrees of freedom may also imply that the expected drop of fit by 1 G^2 unit per parameter optimised need not apply exactly in the asymptotic case (in cases of finite samples, sparseness is probably the more serious concern, although the interaction of these two factors has yet to be ascertained).

For the example in figure 6.1 there are many cells (256), but the number of parameters being optimized (28) is also relatively large. Accordingly, the effect of partial recovery of degrees of freedom on the asymptotic distribution is difficult to judge. Asymptotically, as a rule of thumb, rejection of this model would be checked against a χ^2 with 255 d.f. So, due to recovery of

degrees of freedom, the test used in figure 6.1 may be closer to full Monte Carlo simulation results than expected.

If rejection (or not) of the observed fit of model to data is the primary aim, a step wise approach is useful in the interest of computational efficiency. After finding an overall optimal model for the data, firstly choose samples from the expected pattern probabilities (e.g. as in figure 6.1). If this rejects the data coming from the model, there is no need to look further. Otherwise, the next step is to take samples (or reuse the earlier samples) and evaluate each ones likelihood on the optimal tree (or perhaps all trees in a 99% confidence set), (and reoptimising all free parameters for each sample e.g. edge lengths, tr / tv ratio etc.). If this still does not reject the model, a search across trees may be implemented, and even more extensively, a selection of the mechanism which gives the best fitting model, using a model selection criterion like that of Akaike (see Sakamoto *et al.* 1986, Miller 1990).

6.2.3 Ways to overcome sparseness distorting asymptotic expectations

It would be most desirable to find reliable ways in which to measure goodness-of-fit without the burden of what will be huge simulations in many typical phylogenetic analyses. Here we consider some possibilities.

A traditional approach is to group cells to bring the cell count up to some minimum level. It is often accepted that if the all cells have at least two observations, sparseness effects are minor if cell counts are even. with sequence data however, the unevenness of expected counts per cell suggests using the more conservative guide of at least 80% of cells having an expected count of 5 and none with fewer than 1 observation expected. If grouping of cells is done randomly, then the power of the goodness-of-fit statistic to detect departures from the model will usually fall (Stuart and Ord 1990, p. 1180). However, if grouping is done in line with likely violations of the model, there is a good chance of increasing the power of the goodness-of-fit statistic by grouping cells likely to be biased in the same direction (either up or down relative to expected size).

One problem that arises when grouping cells is that the grouping for the X^2 or G^2 statistic should be done on the basis of expected and not observed cell counts. This immediately presents two problems. One is that most ML methods will very quickly (as t increases) find it prohibitively expensive to evaluate all these patterns (although an understanding of expected parallelisms and convergences could mean that the probability of a large proportion of all expected patterns could evaluated very quickly as an approximate approach). The second problem will dog even Hadamard conjugations predicting all expected pattern probabilities; this is the recommended grouping of cells could be quite different under different trees. This problem is aesthetically unpleasing, but some studies indicate it may have little effect of the validity of any test (Stuart and Ord 1990, p. 1172). One possible way to circumvent this problem would be to use the Neym n statistic, which is just the X^2 statistic with the observed and not the expected frequency as the denominator (i.e. Σ(obs. - expt.) / obs.). This statistic is also a member of the power divergence family (Read and Cressie 1988), but is known to have generally lower power

than the X^2 or G^2 statistics. We are not certain if this approach is theoretically sanctioned, or how effective it might be in practice, although it is worth considering.

Another approach is to evaluate the expected value of a goodness-of-fit statistic using the known multinomial sampling distribution of the data. To do this exactly very quickly becomes computationally prohibitive as t increases. A more useful approach is to approximate each cells contribution to the overall statistic with a Poisson distribution (then evaluate the effect of this variable upon the terms of G^2 or X^2 cell by cell, and sum this overall) (it is also possible to limit this evaluation to the smaller cells, e.g. $\mathbf{f}(T)_i$ with expected values of less than 5). As long as the sequences are long, these small count cells will have negligible covariance which can be ignored in the summation. The covariances between the large count cells can be considerable. In these cases it should probably be taken into account, and one possible way to then get an estimate of the overall expected value is to add the mean and variance of the few large cells (which can be treated as approximately χ^2) to the mean and variance of the many small cells. This is beyond the scope of this thesis, but we hope to evaluate it in the near future. Unfortunately even this approach is going to quickly become impractical with standard ML methods. An alternative might be to apply a statistic like that of Kishino and Hasegawa (1989) which measures the expected difference and standard deviation between the likelihoods per cell of two trees, assuming that the observed differences per cell behave as a normally distributed random variable (some recent simulations by Hasegawa and Kishino 1993 suggest these are good approximations, but much more extensive evaluations, especially with more taxa are required).

Exactly what the expected distribution of the difference in goodness-of-fit will be when there is sparseness in the data, is uncertain. Our earlier approximation, only gives an estimate of the expected fit of random data with no parameters optimised to it (e.g. like that in figure 6.1). We suspect that each additional parameter in the fitted model will drop the overall G^2 fit statistic by less than 1. This is intuitively reasonable, i.e. if the fit of G^2 on the data with no parameters optimised is only about 1/3 as large as its expected asymptotic value, then one would imagine that a model with say x / 2 parameters optimised won't necessarily fit perfectly (where x is the total number of cells). Monte Carlo simulations (or possibly exact calculations, see Stuart and Ord 1990, p. 1169) are required to determine by how much G^2 actually drops (our own interpretation of simulations in Reeves 1992 and Goldman 1993a suggest the drop might frequently be of the order of about 0.5 G^2 units per parameter optimised with ML, given few taxa and current models).

A recent study by Gaut and Lewis (1995) appears to show that the G^2 statistic (measured between trees) does not converge to a χ^2 statistic but rather the tail of this distribution falls away too rapidly. We conjectured to one of the authors (P. Lewis) that this is might be due to sparseness (or a multinomial sampling bias), and increasing the sequence length to an appropriate amount (e.g. 10^6 nucleotides) could eliminate this possibility. The alternatives are an effect due to tree selection, due to partial recovery of degrees of freedom with ML estimators, or a combination of these factors.

6.2.4 Overall goodness-of-fit statistics not necessarily reliable

In no area of modeling should an overall goodness-of-fit statistic be used to pass final judgment of whether a model is adequate for the purpose (Miller 1990). Firstly, it is quite obvious that the real process of evolution is not a simple i.i.d. process and it does not take very long to verify this in any real sequences (e.g. fluctuating base composition along the sequence, correlations between changes at adjacent sites). Given enough data (longer sequences and more taxa) any i.i.d. model will be shown up as ill fitting. Additionally, if the data are not i.i.d. this means the G^2 or X^2 statistics are not necessarily uniformly sensitive to the true randomness of the evolutionary process, and could give misleading guides as to which i.i.d. models were really the better approximations.

What most biologists want is to have a test (or more likely a series of tests) to allow them to evaluate when inconsistency of tree selection may be occurring. This is a much harder question to answer, because it is only loosely related to the absolute *goodness-of-fit*. Indeed examples in section 5.3.7 illustrate this. There overall fit is adequate, the difference in fit between the optimal and the suboptimal trees is significant, but the selected tree is probably wrong due to systematic errors.

Overall goodness-of-fit can be useful when there is already confidence that the data evolved by a similar mechanism to that presented by the model. An example is with recently diverged sequences where there has been relatively little time for multiple hits, and even less time to expect substantial shifts in functional constraints (which may accompany covarion evolutionary processes). However, amongst recently diverged sequences, relying upon goodness-of-fit solely may lead to overkill. It seems likely that the clock constrained homogeneous Kimura 2ST model used in section 5.4.3 is perfectly adequate to suggest that there are too many CG and HG patterns to be explained away as multiple substitutions. It could well be shown that other models fit significantly better (certainly as more data is added), but the improvement in fit might largely come from explaining the multiple hits amongst the longest edges in the tree (especially to the outgroup) and not amongst the closely related taxa of most interest. Goodness of fit, does however seem to come into its own in helping to estimate continuous parameters associated with tree models (e.g. the tr/tv ratio, the distribution of rates across sites, or edge lengths, see section 5.3.9).

It has a fundamental requirement that we seek to understand in detail the actual processes occurring in the real data, and gain an understanding of their overall interaction before real confidence can be expressed in any method of tree estimation. Towards this goal it is useful to have some techniques which allow us to "see" the magnitude of data to model errors with different assumptions. It is interesting that one goodness-of-fit statistic under non-sparse cell counts can facilitate identifying particularly aberrant patterns in the data (with respect to the model). This is the X^2 statistic since its size per cell is an immediate guide to fit cell by cell. Its use as a tool in diagnosing departures from the model should not be ignored.

An important point to take from this section is that phylogenetics should not become wed to an endless search for better fitting models. There needs to be a substantial input also from biology, and studies of what types of errors are expected and need most to be controlled. Identifying such errors is addressed later. Then it may turn out that simpler fitting models, like the normal distribution approximation, have great utility.

6.2.5 Some aspects of overall fit of data to $\hat{\gamma}$

To conclude this section we will discuss measuring the fit of data to model when using Hadamard conjugations. As already shown in earlier sections (e.g. 4.7, 4.8, 5.2), the vector $\hat{\gamma}$ is not a statistically efficient estimator of edge lengths under a tree model of evolution. This in turn means that if edge lengths are estimated directly from $\hat{\gamma}$ without taking account of the full error structure, standard fit measures will not be distributed as χ^2 asymptotically (as $c \to \infty$). Thus, the fit of $c\mathbf{s}(T_n)$ (where T_n is the optimal tree picked by a measure such as closest tree) to the observed data, $c\hat{\mathbf{s}}$, measured by a statistic such as X^2 has an unknown distribution. All we can be sure of is that the expected value is asymptotically greater than χ^2 with d.f. equal to $x - p - 1$ (where x is the number of distinct sequence patterns and p is the number of optimised parameters). Thus the distribution of any fit statistic between $c\mathbf{s}(T_n)$ and $c\hat{\mathbf{s}}$ needs to be evaluated by simulations (evaluations which have yet to be made).

Unfortunately this issue of statistical inefficiency undermines the tests made by Lockhart *et al.* (1992) where they had implicitly assumed that $\hat{\gamma}$ was an efficient estimator, and that their X^2 statistic asymptotically had a χ^2 distribution when the model held. As discussed later, similar tests can be made exact but require either an efficient estimator of model parameters or simulations, or perhaps both (as discussed already with the Monte Carlo simulations).

The GLS estimate of edge lengths from $\hat{\gamma}$ is an statistically efficient estimator (given enough data to approach multivariate normality), and both the residual SS or fit of $c\mathbf{s}(T)$ to $c\hat{\mathbf{s}}$ measured by X^2 or G^2 should asymptotically ($c \to \infty$) have a χ^2 distribution with d.f. = $x - p - 1$. Other statistics such as WLS SS will have an unknown distribution when the edge lengths of the tree are not estimated efficiently (although it is possible to estimate the distribution of these statistics under asymptotic conditions, e.g. see Bulmer 1991 for an example with WLS on distances).

6.2.6 Guides to selecting a well fitting model

The general approach is to only add a parameter into a model if it increases the goodness-of-fit by an amount greater than expected due to chance. For a statistic with a χ^2 asymptotic distribution the expected drop is 2 units even for a variable with no relevance to the true process (i.e. the parameter is only 'explaining' random noise). Akaike has shown, using information theory, that this expected increase in fit can be used (asymptotically) to pick amongst models, even if they are not nested (see Sakamoto *et al.* 1986, Miller 1990). The Akaike Information Criterion (AIC) is usually applied in specific reference to ML methods, and the model selected as

'best' is that which maximises lnL - p (where p is the number of fitted parameters, which is equivalent to picking the best fitting model by $G^2 + 2p$). Others have suggested similar statistics, and also modifications when using finite data (see Miller 1990).

To be more confident that an added parameter has dropped the fit by a significant amount it is possible to make a test, since the expected marginal distribution of a single random variable is asymptotically χ^2 with 1.d.f. Thus the probability of a random variable dropping the goodness-of-fit by more than 3.84 is just 5%. A similar test can be made for a collection of parameters. Looking back to the examples in section 5.3.7 figure 5.6, such a test is: do the extra 2t - 3 = 5 parameters of the nonhomogeneous Kimura 3ST Γ model improve fit significantly over the Kimura 2ST Γ model? The actual difference is 55.58 - 37.37 = 18.21 G^2 units, while a χ^2 variable with 5 d.f. is expected to be greater than this number with probability only 0.003. The ratio of tr / tv1 and tv2 is expected to be different amongst such ancient sequences, but here the significance of the test might equally be due to these parameters being able to help explain patterns due to other processes (including base composition drift) since the adequacy of the model must be open to question. With finite data, these tests may be unreliable.

6.2.7 Modifying Monte Carlo simulations to avoid possible parameter biases

Both Reeves (1992) and Goldman (1993a) suggested beginning their simulations from the parameters estimates of the best tree model inferred by ML. This however treats the ML parameter estimates as a known quantities, when in reality they too are random variables. Thus we suggest it is worth exploring a modified Monte Carlo simulation: a bootstrap replicate of the observed data is taken: an ML estimate of all parameters in the model is made for this bootstrap sample: these parameter estimates are used to generate a random sample of the same size as the original data but under the parameterised model: an ML estimates of the optimal parameters is made of this second level sample and all numbers of interest are kept. The first cycle is complete, and another bootstrap sample is taken and so on n times. Each bootstrap estimate finally yields just one set of parameter values. Here the bootstrap is being used to explore the curvature of the likelihood surface of the original data (in a sense similar to Felsenstein 1992), and should improve upon the standard Monte Carlo simulation if there is any asymmetry in the ML surface in the region of the original sample (for example the boundaries formed between trees). Unfortunately, the method may be biased if the bootstrapping of samples (which increases the variance over the original sampling) in some way introduces a bias to the end result.

We have begun to run evaluations on this sort of method but instead of using ML, are taking a random sample from an **f** vector, estimating $\hat{\mathbf{s}}$ then $\hat{\gamma}$, then choosing a tree using compatibility directly from $\hat{\gamma}$. This tree $\hat{\gamma}(T)$ is then used to infer $\mathbf{s}(\hat{\gamma}(T))$. A second level sample of the same size as the original one is taken from $\mathbf{s}(\hat{\gamma}(T))$, then $\hat{\gamma}$ is estimated from this sample. From this last $\hat{\gamma}$, the parameters of interest are taken (after tree selection again by compatibility). Contrary to our expectations, the variance of a key statistic, namely the sum of absolute deviations from $\hat{\gamma}$ (and our overall measure of fit of model to data) did not increase but rather decreased with the

addition of the bootstrapped first phase. This surprising finding is being studied further, and a programming error has yet to be ruled out.

We will finish this section on overall testing of fit of data to model with a comment on the use of the bootstrap to gauge the distribution of a goodness-of-fit measure. Instead of running a Monte Carlo simulation, it is possible to gauge the sampling distribution of the lnLR statistic using the a bootstrap procedure. This involves taking bootstrap samples of the original data, and for each sample measuring the likelihood of the unconstrained model on each sample ($\ln L_U = \sum \hat{f}'_i \ln(\hat{f}_i / c)$ where the ' indicates the bootstrap sample data), then building trees under a certain likelihood model and in addition to keeping the trees for the usual bootstrap comparisons, also keeping the $\ln L_T$ statistic of each tree. The cycle is then repeated. The distribution of G^2 statistic is estimated from the distribution of $v = 2(\ln L_U - \ln L_T)$(for the n bootstrap samples). This may give an overestimate of the mean of the distribution of the G^2 statistic under a specific model, but probably a reasonable estimate of its variance. The advantage of this statistic is that while it is not as precise as that obtained by a Monte Carlo simulation (for a specific model), it is probably more useful in practice since it costs practically nothing extra to calculate if already doing a bootstrap analysis (Felsenstein 1985). It should be possible to evaluate this statistic for many taxa (e.g. PAUP* can evaluate a 30 taxon tree via simpler likelihood models in under an hour on a fast PC, so should allow overnight bootstrapping on more powerful machines). Paired differences (i.e from each bootstrap sample) between two different ML estimates assuming different mechanisms of change can also be used to test the validity of adding in extra parameters. If the difference in G^2 of one model relative to another is greater in 95% or more of the paired samples (estimated by this bootstrap routine, and without reoptimisation of model parameters), it seems likely that there is a significant difference in the fit of these two models (that is, one fits better beyond just sampling error).

6.3 TESTS OF THE GENERAL SUITABILITY OF A PHYLOGENETIC METHOD

Here, new tests are described which more directly evaluate specific departures from expectation that may indicate inconsistency of tree selection, rather than just a reduced ability to predict the exact frequencies of observed site pattern frequencies. These are conceptualyl based on suboptimal trees being drawn from a particular distribution, which asymptotically becomes a mixture of non-central chi-square distributions.

6.3.1 The expectation of equally well-fitting suboptimal trees

This first test is immediately suitable for four taxa, but can be extended to more. It is based upon the expectation that under a true Markov tree model of evolution, all trees which do not share any internal partition with the true tree will in expectation be equally badly fitting. For finite samples, asymptotically (and under i.i.d. mechanisms) this expectation becomes: all trees which do not share any internal partition with the true tree will be distributed as a noncentral χ^2 variable (with degrees of freedom equal to approximately $x - p - 1$, where x is the total number of distinct patterns, and p is the number of statistically efficiently estimated parameters). This

means that all trees which do not share any internal edge in common with the true tree should be expected to be equal in overall fit to the model, and in actuality only different within sampling error. This expectation is intuitive when considering the selection of a tree in a multivariate normal space which has been made orthogonal (so that selection of any edge in the tree cannot improve 'fit' if it is not in the true tree).

That all trees with no partitions in common with the true tree should all fit about equally well leads to a number of useful results. Firstly, a simple test of the adequacy of the model for tree selection with four taxa is that the second and third best trees should be identical in fit bar sampling error. Thus, if we were to sample repeatedly from the true data distribution, the second best tree would be better than the third best tree 50% of the time with an unbiased method. A practical test of this expectation can be made with the bootstrap: if the second best tree is better than the third best tree more than 1 - $\alpha/2$ of the time, reject the hypothesis of equivalent fit, which suggests evidence for a systematic bias in the data which is directly affecting the fidelity of tree selection. The level α would usually be set at a value near 95% for a single test. It is also appropriate that this test is two-sided since there will usually be no strong a priori anticipation of which of the two non-optimal trees will be better.

The bootstrap can be computationally expensive, and here the tests of Kishino and Hasegawa (1989), or Kishino *et al.* (1990) may act as a substitute. Note that while the Kishino-Hasegawa test compares the likelihoods of two trees and we spoke earlier in this section of the expected differences in G^2, the test is still valid since in this application the G^2 of two differs by just their likelihoods (and the constant factor 2). Our test may also be used to test any two trees without internal edges in common with the optimal tree for any number of taxa. However, it is difficult to decide which two trees to randomly pick. A similar test can be made of parsimony tree lengths (i.e. is one of the two suboptimal trees significantly better fitting that the other), and again Kishino and Hasegawa (see also Templeton 1983) offer a useful approximation to bootstrapping.

An interesting area in which to apply these tests is to very anciently diverged molecules with four taxa, where either Felsenstein or anti-Felsenstein zone effects could be distorting the reliability of tree selection under simple models. In either case, over much of the range of branch lengths (x in sections 5.5 and 5.6) the effect is detectable as the second and third from optimal trees fit quite differently (at the inconsistency point in the Felsenstein zone, two trees fit equal well with a third markedly worse, while in the anti-Felsenstein zone at the inconsistency point all three trees fit equally, but then as internal edge lengths tend to zero, there is the tendency for just the true tree to fit worst). One area where we have already claimed possible inconsistency is in the analysis of four ancient rRNA sequences. This test does not detect any unequalness of the second and third best trees in either the data for figures 2.7 and 5.1, or in figures 5.6a-c. Here there appears to be very little direct evidence of any inequality of the two non-optimal trees, yet as already discussed there is good evidence to suggest the favoured 'eocyte' tree is incorrect on this data (see chapter 3, sections 3.7 and 3.8). The reason for the test of the two non-optimal failing to discriminate in this case, may be partly because this a possible anti-Felsenstein zone

problem, and both non-optimal ML trees have internal edges very close to zero in lengths (and bounded below by this value), allowing little room for their differentiation.

An analysis of ancient functional sequences which seems ideal for this new test is that of Golding and Gupta (1995), who claim a chimeric origin for the eukaryotic genome. If their result is purely artifactual due to strong biases making tree selection unreliable, we would expect a test of the difference in fit of the second and third from optimal trees (with four taxa) to show some evidence (especially since they analyse 24 different genes). We do not have the original data to make this test, but Gupta and Golding's table 1 does give a list of standard errors estimated using the methods of Kishino and Hasegawa (1989) between the optimal tree and the two nonoptimal trees. Assuming the fit of the two nonoptimal trees to be uncorrelated, with respect to the difference in fit from the optimal tree, we make the test. The results are: Using amino acid ML to evaluate trees, two genes showed clear and strong evidence of violating the expected equality of the second and third best trees, these being aminotransferase and beta-galactosidase (both of these being two amongst the 11 genes identified by Golding and Gupta as being able significantly resolve the optimal tree). In addition, two other genes are slightly suspect (deoxyribodipyrimidine photolyase and pyroline-5-carboxylate reductase), and it would be good to test these with the bootstrap itself (which we suspect may be a more powerful test than the Kishino and Hasegawa (1989) approximation).

Similar tests applied to parsimony on the same genes yield quite different results, and none of the genes fail the test, despite 17 showing significant support for the optimal tree over both alternative arrangements (there are just two genes which are slightly suspicious, these being carbamoylphosphate synthase and anthranilate phosphoribosyltransferase). One the face of it, these results seem to back up Gupta and Golding's (1995) conclusions. However, the differences between ML and parsimony deserve further study, with one possibility being that the ML model is giving spurious results for some proteins. The biological question which needs further work before accepting Gupta and Golding's claim of the eukaryotic nucleus arising from a fusion, is to confirm that the genes supporting a eukaryote-gram negative tree are not explainable by migration to the nucleus of mitochondrial genes. This will require sequencing of homologous genes in the amitochondrial eukaryotic taxa, especially Microsporidia and Diplomonadida.

6.3.2 The general distribution of the likelihoods of trees under a reliable model

Noticing that all trees with no internal partition in common with the optimal tree will be have G^2 distributed as a non-central χ^2 variable under the true model has some other useful applications. For example, all trees which share no partition with the true tree should asymptotically have G^2 statistics that differ approximately as variables drawn from a χ^2 distribution with d.f. equal to approximately 2(p - edges in common) (where p is the number of edges in the tree). A chi-squared distribution quickly becomes bell shaped as it converges to the normal distribution with increasing degrees of freedom. Thus for any reliable phylogenetic data evaluated under a matched likelihood model, the true distribution of G^2 will have a predominantly bell shaped distribution with a skew of generally better (lower) G^2 values

corresponding to the trees with some correct internal edges extending to the left. To the extent that the best tree is clearly separate from the bell portion of the curve and also all other trees, then the data set is resolving all internal edges. If, say, one internal edge is well resolved, then there should be a distinct bimodal distribution (e.g. Penny and Hendy 1987 observed this). With two internal edges very well resolved there could be three modes. As far as I am aware, this expected distibution of tree fit has not be explained as an asymptotic statistical result.

Thus a possible test of goodness-of-fit test to assess the data clear signs of inconsistency in tree selection, is to ask if the trees which do not contain any edges in common with the optimal tree conform to expectations. We presently working on exactly what these will be in general. One way to make a slightly conservative test is to get some idea of the expected distribution using Monte Carlo simulation. We suspect that the distribution of G^2 between the model predicted data and a random sample (without even selecting a tree) should be useful. Once this distribution is measured accurately (say 10,000 simulations), it can be compared with the observed spread of tree fits by first rescaling them so that the mean of both distributions has the same value. Then apply a test for any sign that the true distribution is more spread out than the expected distribution. A Kolmogorov-Smirnov type of test should be useful (with modification since both the observed and the expected distributions are estimated from finite samples). This test may suggest bias quite often with real data, but should be more to the point of assessing reliability for tree selection than the overall goodness-of-fit test.

The expected theoretical fit of G^2 to all trees also helps to explain the observation that tree lengths under parsimony and distance criteria tend to be more asymmetric the more the data resolves the tree (e.g. Archie 1989a, Hillis 1991). By comparing the size and significance of a skew in tree lengths it is possible to crudely gauge whether the data set has distinct hierarchical structure (not necessarily phylogenetic, as it could also be due to strong base composition biases for example). It is dangerous to say that a data set has phylogenetic structure, just because it is skewed. For example, a data set with strong base composition biases and a proportion of invariant sites can appear to have significant structure even when variable sites are completely random. This can hold for both within column randomisations, row randomisations, or both (unless the invariant sites are first removed; Peter Lockhart per. comm. is also aware of this and tends to make tests like those in Steel *et al.* 1993, after first removing all constant columns). With a continuous distribution of rates across sites, the problem is even less tractable, since a base composition shift lag (see chapter 3) makes it difficult to separate out hierarchical structure due to phylogeny alone. In this situation dropping all nonparsimony informative sites helps, but may not cure the problem.

6.3.3 Extensions to split decomposition

For sets of more than four taxa, a useful way to visualise biases of the nature that the second and third best trees have quite unequal support is to use the method of split decomposition (see Bandelt and Dress 1992). Although to date this method has started with the difference in sums of pairwise distances, it could also start with quartet data based on four taxon trees evaluated by ML for example. One possibility would be to give the best tree for any four taxa a score of 2, and

the second best tree a score of 1. Alternatively the length of each quartet could be estimated as the difference in the sum of edge lengths of the best tree minus the worst binary three taxon tree, and the length of edges on the second best minus those of the worst tree. Another possibility might be the difference in log likelihood score. In any case, the aim would be to look for consistent patterns in the data, possibly due to systematic biases. Another quartet method which might work well in combination with splits graph drawings might be evolutionary parsimony invariants (Lake 1987), or the linear invariants now known to exist for the Jukes Cantor model (Fu and Steel 1995 and section 4.6.2). Note that at present the method of split decomposition does not differentiate random error from systematic bias, although by marrying the method to bootstrapping this could be achieved (that is assign to all features of the original graph the frequency with which they appear in graphs estimated from bootstrap samples of the original data). Methods such as the distance Hadamard may also offer potential for visualizing patterns not showing up in the optimal ML tree.

6.3.4 A sign test for the fit of $\hat{\gamma}$ to model expectations

Asymptotically as $\hat{\gamma}$ tends to having a multivariate normal distribution, the expectation is that apart from the entries $\hat{\gamma}_0$ and those corresponding to edges in the tree, half of all entries should be positive and half negative if the data fits the model. An appropriate (but not optimally powerful) asymptotic test of fit of data to model would be a binomial test that the observed proportion of signs on these entries in $\hat{\gamma}$ was 50:50. A one tailed test can also be appropriate; for example if an i.r. correction is made, but the anticipation is that there will be too many positive entries in $\hat{\gamma}$ (due to undercorrection for multiple hits). Application of this test to the transformed data in figure 2.9 gives a very clear answer: there are 64-15-1-1 = 47 non-tree patterns which are all positive and the binomial probability of this is just 1 in 2^{47}, or about 10^{-14}!

Unfortunately, sparseness of the data will distort this binomial type of test, since there will be many entries in $\hat{s}$ with value zero, and these which will tend to translate to slightly negative entries in $\hat{\gamma}$ (due to the binomial like marginal distributions of these entries with expectation near zero, see section 4.3). Consequently, a one tailed test of too many positive entries in $\hat{\gamma}$ will tend to be conservative, especially as the number of taxa grows (and especially when using 4-state data). A possible modification would be to filter out all the values within some small amount of zero. If this ε is chosen appropriately (a value $\varepsilon = 2/c$ would remove much of the exponential character of marginal distributions) and would greatly reduce the conservativeness of the test. The properties of tests need further study, and verification that this conditioning on larger values is not causing a bias in the test to be significant too often.

An alternative approach may be to test pathset lengths with a sign test. Recall that every higher order pathset length can be estimated as a sum of observed distances (usually by the minimal sum criterion, but also able to be read directly from the optimal tree). If a correction formula is undercorrecting on a higher order pathset, then we expect that the sum of inferred pairwise distances homologous to this pathset will be larger (since the distances will each be less

than the pathset length and so incur fewer multiple substitutions). One factor pushing the test towards being conservative would be the minimality criterion in estimating pathset lengths from distances (most strong when the tree has many short internal edges). A factor operating in the opposite direction will be the nonindependence of paths and pathsets, which could result in too many rejections of the null hypothesis if the final test was based on a binomial distribution. Simulations could be helpful to determine reliability of this test given these two opposing biases.

6.4 COMMENTS ON THE BOOTSTRAP

Bootstrap resampling is widely used in phylogenetics (Felsenstein 1985, Penny and Hendy 1985), particularly in the form described by Felsenstein where it is used to estimate the reliability of individual internal edges in the tree (i.e. the probability that this edge exists in the true tree). A drawback of the method is that it has been shown to be particularly biased towards underestimating the reliabilitly of an edge in the true tree (Zharkikh and Li 1992a,1992b, and Hillis and Bull 1993) if the support is large (generally > 50%). On the other hand if the support is low, then the bootstrap tends to give a too high a probability estimate of the reliability of an edge in the tree.

6.4.1 Approximately estimating the bias in bootstrap support for edges in a tree

The aim here is to evaluate how much bias the bootstrap procedure may be causing in estimating the support for a particular edge in a tree. This work is an offshoot of the earlier analysis to modify Monte Carlo goodness-of-fit simulations to accommodate uncertainty in the original parameters. The proposed method to estimate bias in bootstrap support for an internal edge is to: take a bootstrap sample (s') and infer a tree from it. Rather than discarding the original sample immediately, another single bootstrap sample (s") is drawn from it and a tree also inferred for this sample. Lastly, a single pseudosample (s"') is drawn from s" and a tree is inferred from it before discarding this series of samples, and repeating the whole process n times. Thus we have three levels of trees, and on each the support (b.s.p.) for an edge in the original tree is made by counting its frequency of appearance (Felsenstein 1985, 1993). Accordingly, this gives three numbers (b.s.p. 1, b.s.p. 2, and b.s.p. 3) for each edge in the best tree from the original data. Next, plot each of these numbers as (1, b.s.p. 1), (2, b.s.p 2) and (3, b.s.p. 3) on a Cartesian graph. The way these numbers change as we repeat our sampling, and so mimic the original bootstrap bias, is our guide to the unseen bootstrap bias from the original sample. In order to infer the adjusted bootstrap support for the original optimal tree, fit a curve (the form $x = ay^2 + by + c$, seems reasonable) through these numbers and take the y intercept as our estimate of the adjusted bootstrap support (subject to the condition that it cannot be lower than the b.s.p. 1 value). Presently, we are not sure how this method relates to the complete and partial bootstrap recently suggested by Zharkikh and Li (1995).

Another intuitive way to overcome the bias in the bootstrap is to increase the size of the resampling over that of the standard size c (just as to account for positively correlated sites you could decrease the number of resamplings to less than c). Unfortunately, even if we could make a reasonable estimate of what the new resample size should be, the effects of these modifications

on bias could be partly unpredictable. It is possible such modifications will overcompensate more for the edges in the tree which are expected to show the least bootstrap bias (generally shallow edges), but undercompensate for the deeper edges in a tree (which tend to have the most downward bias).

One factor found already in our simulations, and reported in Waddell *et al.* (1994), is that the delta method provides nearly unbiased estimates of the variance of entries in $\hat{\gamma}$ when working from just samples, but underestimates when working from the original **s**(T) vector. The opposite occurs with resampling with replacement. When sampling with replacement is performed upon **s**(T), under the assumption of independent sites, it gives the exact variance of entries in $\hat{\gamma}$, but an overestimate when applied to each sample, $\hat{\mathbf{s}}$, in turn (i.e. the bootstrap). Part of the reason for this is that the bootstrap is a nearly unbiased estimate of the variance in **s**(T), but the nonlinear pathset length transform causes an additional bias as already discussed in chapter 3. This added effect which increases the bias of the bootstrap to underestimate edge support for transformation based methods is evaluated in more detail in section 4.8.2.

6.4.2 The number of alternative trees and bootstrap bias

With many alternative trees bootstrap estimated reliabilities may be quite conservative (Zharkikh and Li 1995). In addition, it is also perceived that bootstrap estimates tend to become more conservative the more deeply within the tree an internal edge is. However, many short internal edges are 'bracketed' between well supported internal edges, and this results in very few likely resolutions of these edges, and consequently much less bootstrap bias in estimating their support. For example, consider the edge defining the archaebacteria which has only three real alternatives because the eukaryotes, eubacteria, methanogenic and eocyte groups all have large bootstrap supports (al greater than 90%, two being practically 100%). Hopefully, we should be able to measure this effect and use it to gauge just how many effective alternatives there may be to an edge in a tree, and so gauge how conservative the bootstrap may be in that instance. In the archaebacteria example, the guess would be slightly more than 3 effective alternatives.

Alternatively, if the edge to be tested is known a priori, constraints can be made to force there to be just three alternatives (the program PAUP*, by Swofford 1993 allows any combination of compatible edges to be forced to be present in the inferred trees). Knowing the number of alternatives, the bias with just three alternatives is amenable to direct calculation based on multinomial, or multivariate normal distributions (e.g. Zharkikh and Li 1995). If there are only three alternatives, bootstrap support needs to be of the order of 85% or greater in order to correspond to a 95% probability that the edge is in the true tree (which is like saying a 100 - 95% or 5% chance that another tree could give data with this much support for the edge in question, Felsenstein and Kishino 1993). This would imply that while some of the six a priori hypotheses in chapter 3 figure 3.11 have strong support (e.g. (IV) plants and fungi closest relatives, (VI) Thermus deep), but two others are marginal. These being (I) Archaebacteria monophyletic, which has dropped below 70% with 20% of constant sites removed, and (II) Microsporidia earliest, which is in the region of 80-90% support with 20-30% of constant sites

removed. A final decision on whether these show sufficient bootstrap support will require evaluation of how much downward bias may be generated by bootstrapping followed by the non-linear LogDet transform (especially with constant sites removed).

An irony of so much attention on the conservatism of the bootstrap under i.i.d. models, is its possible overestimation of support with real molecules with strongly correlated site changes. A good example being the paired stem regions of rRNA and tRNA's. In some of these molecules' sites, substitutions almost always occur in pairs, e.g. if a C → T, then on the complementary strand a G → A if the original substitution is not reversed or lost from the population due to selective disadvantage. For the actual tree reconstruction process it may well be appropriate to weight such a change highly if it is very rare, however for the purpose of assessing statistical support it should be treated as a single change, which means that when one of these two sites is picked in a bootstrap resampling, the other one is automatically picked as well. Unfortunately, such modified bootstrapping does not appear in the major programs, and without good studies of just which sites have such high correlations, may be of little practical incentive to use. More of a challenge will be to incorporate more subtle, but still influential correlated changes, and their effect upon statistical support (section 4.9 where the variance covariance matrix may be used for similar purposes).

Yet another approach to using the bootstrap, is to evaluate a 95% confidence interval on how many edges in the optimal tree are likely to be wrong. This can be done by a similar approach to Penny and Hendy (1985) except rather than taking the mean value, find the partition metric distance of Foulds and Robinson (1981) below which 95% of inter-tree distances occur. This will hopefully give a realistic estimate of how many edges (y) may be wrong in the whole tree. This statistic may be useful to test that the data definitely do have hierarchical structure. Such a test is probably more appropriate to intraspecific data like that of Vigilant *et al.* (1990), which is expected to do poorly on the standard bootstrap approach because single edges are typically only supported by one or two substitutions, despite good evidence that hierarchical structure is present (e.g. Penny *et al.* 1995). In such situations, keeping the y best supported edges by the Felsenstein (1985) measure of bootstrap support per edge, may give a useful visualization of the extent, if not necessarily the exact location, of edges likely to be correct.

6.4.3 Subtree extraction to counter conservatism when adding extra taxa

Yet another way to use the bootstrap more effectively in a sampling situation, occurs when we are really interested in the relationships of particular sequences, but have added related sequences in the hope of stabilising the tree against systematic and / or possibly stochastic errors. An example might be to test the relationships of Microsporidia to diplomonads to, trichomonads to mitochondrial eukaryotes using 16S-like rRNA molecules. It is often desirable to have more than one outgroup sequence to counter systematic errors, and adding such additional sequences is unlikely to cause the problem of prokaryotes being mixed up with eukaryotes due to sampling errors. However, it is just as desirable to add in deeply divergent Microsporidia, etc. for the same sorts of reasons. Because these sequences are deep, and internal edges may be getting short, there

is an increased chance of sampling errors putting them in the wrong place on the tree. This need not disturb our conclusion since there is already confidence that Microsporidia (and the other groups) are each monophyletic (in the sense of Hennig 1967). An appropriate way around this problem would be to nominate just one sequence from each group as a 'type' sequence to represent that group (preferably the 'type' will be on a lineage with evenly spaced internal nodes leading to other taxa). After the bootstrap is performed, all but the 'type' sequences are removed from the trees and the frequency of what are now, say, edges in five-taxon trees is calculated. Here the bootstrap is much less likely to be overly conservative. The results of this 'subtree extraction' approach will be interesting to compare with the similar but distinct enforced constraints approach mentioned in the previous paragraph.

This subtree extraction approach could also be extended by giving each member of a group a membership score rather than nominating just one 'type' sequence. The higher this score, the more a sequence is considered representative of that group. When is comes time to collate the results of many bootstrap trees of all sequences, a decision must be made to define one subset of sequences in each tree as being representative of that group's position. For example, there may be five Microsporidia species of which four cluster. If these four species have a membership score higher than the fifth sequence their position (or rather that of their last common ancestor) relative to the other groups is taken as representative for just that tree. This type of approach aims to weight against a single representative sequence or pair of sequences being misplaced.

This is the end of these comments on the bootstrap. Our aim is to implement them as soon as possible, and study further their effects. I am presently uncertain how much part of this work overlaps with that of Zharkikh and Li (1995), which needs to be clarified.

6.5 TESTING FOR SPECIFIC DEPARTURES FROM THE MODEL

It is often more desirable to test for specific departures from the model rather than to simply test overall fit of data to model. Such tests can be more powerful at detecting departures from the model, and are also diagnostic in that they can suggest where the model could be modified in order to best improve its robustness.

We begin by looking at testing individual entries in $\hat{\gamma}$, and show how $\hat{\gamma}_{/sc}$ can be a highly informative way of presenting the effects of different transformations upon the data. Following this, specific types of test at the level of the observed data are described: many tests at the $\hat{s}$ level are likely to be more powerful and simpler to perform than equivalent tests at the level of $\hat{\gamma}$. The reason for the increased power is that simple statistics like X^2 applied at the sequence level, take i.i.d. model expected variances and covariances fully into account, whereas doing this at the $\hat{\gamma}$ level requires much more computation, and assumptions of multivariate normality.

6.5.1 The fit of individual entries in $\hat{\gamma}$

As shown earlier in section 4.3, if the variance of $\hat{\gamma}_i$ is greater than $(c\text{-}5)(5)/c^2$ then it has an approximately normal marginal distribution. This in turn allows a simple univariate normal test

of whether entry $\hat{\gamma}_i$ has expected value zero (using the variance estimates from section 4.2). Dividing each entry in $\hat{\gamma}_i$ with variance greater than $(c-5)(5)/c^2$, by the square root of this variance gives $\hat{\gamma}_{/se}$ ($\hat{\gamma}_i$ divided by its standard deviation), and each such entry can then be treated as normally distributed with standard deviation = 1.

Figure 6.2 shows an example of $\hat{\gamma}_{/se}$ for the data in figure 5.1, and with varying proportions of invariant sites removed. The data presented in this way clearly shows how the signal-to-noise ratio decreases the more invariant sites are removed. In this case the relative support of the size of the external edges remains in the same rank as their size. Figure 6.1b also clearly shows that with all sites assumed variable the model fits badly, but with about 28% of sites assumed invariant the data fits nearly perfectly to a tree model. However at this point the most support for an internal edge (in this case the eocyte tree) has dropped to only 0.7 *z* units, which is not significant (this result agrees well with what ML tree fitting to this data suggests in section 5.9). The other interesting feature is that again the existence of a double minimum for the fit of data to model assessed by WLS is evident. The first minimum is at approximately 0.285 and the other at about 0.375. At the second minimum, edges tend to either positive or negative infinity but their standard errors tend to positive infinity more quickly.

Figure 6.2 contains the information to make single tests of all entries in $\hat{\gamma}$, e.g. is the support for the halobacterial tree significant at the 95% level with all sites assumed variable? This potential internal edge, shown in green in figure 6.2, initially has 2.25 *z* units of support and specifying a one sided test we reject that the expected value is zero or less at the $\alpha = 5\%$ level. Multiple tests become more difficult, but a series of squared normal deviates with near zero correlations can be treated as an approximate χ^2 random variable. For example, with 20% of invariant sites removed is the support for the potential internal edges supporting all three trees greater than zero (if so, the data fits the model and a star tree cannot be rejected). The sum of squares of the relevant $\hat{\gamma}$ values at this point is $0.25^2 + 0.75^2 + 1.2^2 = 2.065$, which is not significant when compared to a χ^2 distribution with 3 d.f. ($p = 0.559$). Thus, the data appears to fit the model but a star tree cannot be rejected. All these tests are probably conservative in practice because there is a strong positive correlation between the realisation of a random variable $\hat{\gamma}_i$ and its estimated variance (see section 4.4),

If the variance of an entry in $\hat{\gamma}$ is less than $(c-5)(5)/c^2$, it may be most appropriate to test assuming it has a binomial like marginal distribution. For example, if entry $\hat{\gamma}_i$ has size 0.001 with $c = 2{,}000$, and an estimated standard deviation of 0.0005, then an approximate test that this entry has expected value zero is: does $(\hat{\gamma}_i + 0.0005) \times c = 3$ fall in the upper 5% range of a binomial random variable with mean $0.0005 \times c = 1$. The answer is no ($p = 0.080$) so as a result we do not reject it having a mean of zero (the opposite conclusion would be reached if the test was made assuming a normal marginal distribution, since the observed value appears to be two standard deviations above the expected mean). In making this test, the standard deviation is being used to

estimate the expected value of the binomial variable (as 0.0005 × c), and the deviation of this binomial random variable from zero is taken to be its observed size plus its expected value i.e. ($\hat{\gamma}_i$ + s.d.[$\hat{\gamma}_i$]) × c. This adjustment needs to be made since entries in $\hat{\gamma}$ have expected value zero.

With many possible entries to test, the problem of multiple tests becomes serious if there are no strong prior expectations. This can be alleviated with a multiple Bonferroni type of correction (e.g. Fleming 1982, Rice 1989, Rom 1990), where the number of entries in $\hat{\gamma}$ minus one could be taken as the number of tests being made. More realistically, under sparseness assumptions, a useful guide to the number of tests being made might be the number of cells in $\hat{s}$ relating to two or more observed site patterns.

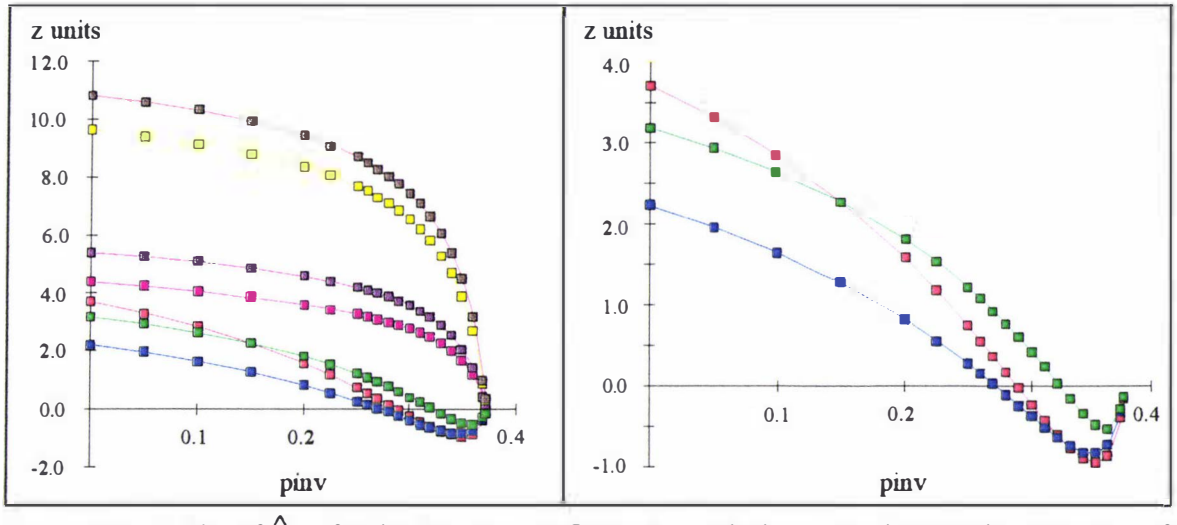


FIGURE **6.2a** A Plot of $\hat{\gamma}_{/se}$ for the data used in figure 5.1, with the x-axis showing the proportion of constant sites treated as invariant. **6.2b** The signal of relating to just the internal edges. The brown symbols show the support for the edge leading to humans, yellow leading to *E. coli*, deep purple leading to *Haloferax*, light purple leading to *Sulfolobus*, red the support for the archaebacterial tree, green the support for the eocyte tree, and blue the support for the halobacterial tree.

These tests of individual entries in $\hat{\gamma}$ are not necessarily a good substitute for the bootstrap method of evaluating the stability of a tree to resampling error. The problem of course is that two or even three directly competing signals could all be significantly greater than zero (as they are when no invariant sites are removed in figure 6.2, and see section 4.8.2 for evaluations of tree selection likelihood which take competing entries into account). This 'univariate' approach to evaluating the reliability of a tree has parallels with other methods of tree estimation. These tests of entries in $\hat{\gamma}$ are very like the confidence intervals Gouy and Li (1989) used to estimate support for internal edges of their OLS distance trees. However, in the estimation of distance trees, because information is lost, it is usually the case that only two of three possible resolutions of an internal edge will show a significant positive length. Observing this is a possible indication that the data is showing a significantly bad fit to the model. With maximum likelihood, the ML length of an internal edge in a tree can be evaluated by comparing it with a zero length.

Here again it is possible for all three resolutions of an internal edge to have significant support by this 'univariate' test. Again it does not show that one tree is better than the others, but rather the possibility of a lack of fit between data and model. These internal edge length type tests tend to do better if they are made between the best tree and the second best tree (e.g. see Gaut and Lewis 1995), but even here the significance level needs to be adjusted if an ironclad commitment was not made a priori to test just one of the three resolutions about a specified internal edge (e.g. see Zharkikh and Li 1995).

6.5.2 Comparing actual and predicted numbers of observed changes per site

This test refines those in sections 3.6.4 and 3.6.7 and predicts the expected values under a specific model. For this test to be performed in its entirety probabilities of all expected sequence patterns need to be calculated under a specified model (tree plus mechanism of character evolution with a statistically efficient optimisation of parameters by a method such as ML). The next step involves summing up the probabilities of all patterns which would have an unweighted parsimony length of 0 on the tree of interest (i.e. the probability of all the constant sites); then the sum of all patterns which show just 1 parsimony change on the tree; then all those with 2 etc. up to the maximum number possible. Compared against these expected numbers are counts of the observed number of sites with 0, 1, 2, etc. unweighted parsimony changes. This gives a contingency table which may then be tested for equality of observed and expected using standard goodness-of-fit statistics like X^2 or G^2.(see for example figure 3.9 and table 3.4). Of course, care should be taken to group adjacent cells (in ascending order perhaps) to ensure that no more than 20% of cells have an expected value of less than 5, and none are less than 1 (and with the degrees of freedom being the final number of cells in the contingency table, minus 1). If the overall fit is not adequate, then it appears possible that there is a mismatch of the model's assumed distribution of rates over sites (e.g. a Γ distribution) and the true distribution (which might for example be bimodal). (And note that of course such a test can only be exactly indicative of the expected problem, if we assume that the rest of the model is correct).

If we have a poor fit to the rates over sites, then it would be useful to diagnose it further. Another possibility is that the model is not assuming enough invariant sites. Fitch and Margoliash (1970) considered this possibility, and tested it by allowing for a fraction of invariant sites. Their test was very similar to the one we propose, but instead of calculating the expected numbers of changes per site exactly, they used an approximation based on a Poisson distribution (see section 3.6.3). Their test is to measure a X^2 statistic before and after allowing for this extra parameter (see table 3.4 for some examples), while we propose the same test, but using the exactly calculated expected numbers of changes. Goldman (1993b) proposed a single cell test of the observed numbers of sites with no changes, against the exactly calculated expected number. Such a test is binomial under the i.i.d. model (e.g. see chapter 4), but because the number will usually be much larger than 40, it is quite reasonable to use a Poisson, Normal, or X^2 (1.d.f.) approximation (reverse order of ease and exactness). All of these tests are made without reoptimising other parameters in the model. As tests in section 5.3.5 and 5.9.3 showed, the binomial test tended to be much more powerful than a likelihood ratio test with reoptimisation of

all the other parameters (e.g. if there really are invariant sites, an ML model will try to minimise their effect as much as possible using the free parameters at hand, and in balance with other parts of the model which do not fit). The most powerful tests for an excess of invariant sites possibly come from full Monte Carlo simulations (e.g. see Reeves 1992a, and Goldman 1993b).

Probably the most sensible use of a X^2 test of observed parsimony changes versus model predicted numbers is exploratory. In particular by looking at the X^2 statistic per cell, it should be possible to get a glimpse of which cells have the worst fit. If they are all the cells with the greatest number of changes, then our model may be missing a proportion of hypervariable sites. If there is a general trend of too few observed changes (for the fewest numbers of changes per site), followed by too many, followed by too few followed by too many, then we may suspect a bimodal distribution of rates across sites. Tests for these features can be described, but their application needs to be wary of the problem of multiple tests, and especially testing after a trend is seen in the data. Just observing substantial fluctuations in the proportions of observed sites in the different rate classes should then be motivation for trying a more flexible model which allows for bimodality, which then allows a test for a significant reduction in the overall fit of data to model for the number of extra parameters.

6.5.3 Testing for an excess of changes predicted on external edges in the tree

Earlier in section 5.3 we saw that a common feature of models which did not allow for a distribution of rates across sites was that they predicted longer external edges than expected. This is done to partly explain an excess of parallel and convergent changes not expected by the model. A test of this feature would be to sum all the unique external edge substitutions (parsimony singletons) expected under the model, versus the total observed number. This could be tested with binomial, Poisson, normal, or X^2 statistics (using the latter three approximations when appropriate (e.g. see Stuart and Ord 1987)). The test would usually be one sided since we would expect an excess of singleton changes predicted by an inadequate i.r. model (of course expectations under a covarion model for example, could be quite different). If something unusual was happening in just one part of the tree, then looking at the individual X^2 cell statistics for observed to expected of each external edges' singleton changes, and comparing them with the groupings of sequences in the tree, could be suggestive of a localised lack of fit.

Lockhart *et al.* (1992), also considered a test of the number of singleton changes per cell against those predicted by $cs(T_n)$ (where T_n was picked by closest tree). They used an X^2 statistic to make the test, then compared the test result with a χ^2 distribution. That particular test is unlikely to be valid since $cs(T_n)$ is not an efficient statistical estimator, so the distribution of the Lockhart *et al.* (1992) test statistic needs to be evaluated by simulations. When a test is made in this context (i.e. with $cs(T_n)$) the original authors did not discuss the diagnostic aim of their test, and it would seem to be two tailed unlike the test of the previous paragraph where we are looking for an excess of singletons predicted by the model.

From this thesis it is clear that if transformations are made under an i.r. model, and there is really a distribution of rates across sites, then external edges estimated directly from $\hat{\gamma}$ will tend

to be too short. Consequently $cs(T_n)$ will tend to predict too few singleton changes. In this case a useful test is to sum all singleton cells, then make a one tailed test for too few singleton changes, as a sign that an i.r. model was not adequate. It might also be predicted that sequencing changes could result in an excess of singleton changes (especially if rates across sites are uneven). However, we conjecture that under an i.r. model, completely random sequencing errors in sequence *i* should have exactly the same effect upon pattern probabilities as making the external edge leading to sequence *i* longer (and this result should hold for all *i* sequences jointly). If this is the case, then there would not be any point interesting for sequencing errors under an i.r. model, but with a distribution of rates across sites it may be detectable (although the best strategy is probably to look directly for unexpected changes along a sequence of known function). A really bad sequence might also show up as a violation of what was otherwise clock like data.

6.5.4 Evaluating numbers of parallel and convergent substitutions

By parallel changes, it is meant that two or more lineages had the same ancestral state, followed by independent changes to the same new state. Parallel changes have long been of particular interest in morphological evolution. In the context of sequence evolution models, Goldman (1993b) suggested a test of excess parallelisms by summing up all the relevant expected probabilities and making an appropriate binomial test of observed against expected. With the more simple i.r. models one might expect too many parallelisms in the real data, and in this case a one tailed test is appropriate. This sort of test could have special use in amino acid coding regions, to test for parallelisms of the sort detected in the lysozymes of primates and birds (Stewart and Wilson 1987). A problem comes with deciding whether a pattern is a parallelism or a convergence (changes to the same state but from different ancestral states), and this can only be answered by attempting to reconstruct the ancestral state. If we take just the bipartition patterns with any number of taxa, it seems likely that these changes were parallelisms. In chapter 5.4.4 tests of this nature were made to help evaluate whether there was a significant excess of patterns in the data consistent with ancestral polymorphism.

6.5.5 The number of states shown at each site

One of the features of the evolution of functional molecules is that some sites may be near neutral in accepting changes, others will be nearly invariant, while others may accept changes but only between certain states (e.g. a pair in rRNA needs to be G:C, or C:G, so changes of C $\leftrightarrow$ G only are allowed). Such sites can be a real problem giving rise to many more parallelisms and convergences than expected under the model. One way to test for this possibility would be to count up the frequency of all observed bipartition patterns, e.g. all A's, A's and C's, ... , A's, C's, G's and T's, and compare this to the expected frequency of all these patterns under the model. An appropriate test would be a one tailed binomial test of there being too many such bipartition patterns. This sort of test may well require a close fitting model; in its one tailed form is should be reasonably conservative if there is an excess of invariant sites (since there will then be too many sites with all the same state, and too many multiple hits generating more than two states at one site).

Another interesting statistic is the number of distinct patterns shown in the data, compared to the expected number under the model. Goldman (1993b) introduced this test, and a full description of it can be found therein. The **s** vector produced by a Hadamard conjugation is ideally suited to this test as it contains all the possible site pattern probabilities which are required to make this test. It might also be a useful test for use with $cs(T_n)$ although it may well reject too often. It would be an interesting statistic to plot as the distribution of rates across sites assumed by the model became more extreme. It may provide a useful statistic to complement overall goodness-of-fit in order to estimate what spread of rates across sites is optimal for a given distribution (e.g. Γ).

6.5.6 Testing for evidence of trapped ancestral polymorphism

A final test of particular patterns which has been found useful is the ancestral polymorphism test given in section 5.4.2.4. It seems important to use this test along with the "control" tests (section 5.4.2.4) to confirm it is not a general feature of the model to underestimate parallel changes.

Overall this set of tests is useful for diagnosing what the major cause for problems under an i.i.d model may be. It is important to also look beyond i.i.d. models. Section 3.6.1 gives a test of a non-i.i.d. model of whether the base composition of unvaried sites is significantly different to that of the varied sites. Another useful test of non-i.i.d features is that proposed by Miyamoto and Fitch (1995), which compares the number of sites seen to be variable in two groups under a Γ distributed rates model and a simple covarion model.

6.5.7 Testing the molecular clock

The basic idea in testing a molecular clock by a goodness-of-fit statistic is that if the constraint of the clock does not worsen fit by a significant amount, then there is no evidence that the clock is being violated. Assuming long sequences and under the true i.i.d. model with independent sites, then the asymptotic marginal distribution of a set of x-1 random variables has expected value $2(x - 1)$, and follows a χ^2 distribution with d.f. x-1. Felsenstein (1988, 1993) and Goldman (1993a) describe such tests and we will not go into any unnecessary detail, except to note again that the clock constraint only makes strict biological sense under a homogeneous model of evolution, but quasi clocks may well be the norm and deserve further investigation regarding their modeling.

Another approach is to observe that under a clock model, the observed F matrix between two groups of species that are strictly monophyletic, should always be the same. Under the Jukes-Cantor model, the efficient statistic becomes the d_{obs} or Hamming distance (since all off diagonal entries in F are expected to be equal). Thus as Tajima (1994) points out, an appropriate test is then just a binomial test that the observed proportions are the same going from one monophyletic group to each member of another group. If the strong correlations that are introduced by shared descent are not taken into account, then there is no easy way to combine the results of more than

one to be sure that the test is not being strongly conservative (reject the null hypothesis to rarely) or too liberal (reject the null hypothesis to often).

The solution to correlations under the multinomial model is to take into account the correlations between patterns, caused by phylogeny, and the easiest way to do this is to use a statistic like G^2 (i.e. to do the model goodness-of-fit test). This test is just the testing of the clock after using ML optimiastion. The likelihood ratio (or an X^2 difference test) can be much more powerful than the pairwise distances test, if there is enough data. There appears to be no straight forward and reliable test of a molecular clock that can be made uniquely in the context of Hadamard conjugations (some thing like GLS fitting $\hat{\gamma}$ may be possible, but is unlikely to be worth the effort compared to ML or minimum X^2). The main concern of the likelihood ratio test is that the data are sparse. A GLS taking account of phylogenetic correlations could be based on observed distances (which should be ultrametric under the clock), although the power is unlikely to be good in comparison to the likelihood ratio or X^2 test if the evaluations of sections 5.2, 5.5, 5.6 and 5.9 are any guide.

6.6 OBTAINING A CONFIDENCE SET OF TREES

An important concept is to evaluate a confidence set of trees. An example would be the that set of trees (in rank from the most to the least supported) which give a cumulative probability (e.g. 95%) of including the true tree when taking account of sampling error. There are two main ways to approach this problem. Two alternatives are: to pick those trees most frequently encountered in resampling the original data (or some approximation to this) versus sampling from data predicted under the most favoured model (a Monte Carlo simulation). In both cases there needs to be confidence in the model in order to believe the resultant set is representative. Members of the first approach include using asymptotic fit statistics (e.g. Bulmer 1991, Navidi *et al.* 1991, Felsenstein 1988), bootstrap resampling (Bulmer 1991, Cao *et al.* 1994, Penny and Hendy 1986) and approximations to bootstrap resampling (Kishino *et al.* 1990, Hasegawa and Kishino 1993, Templeton 1983) to estimate the trees likely to be picked as optimal when sampling error in the original data is taken into account. The second approach has so far only been used by Goldman (1993a); it is very computationally expensive and there is concern over how much it may condition on certain trees and mechanisms of evolution. Little study has been made comparing these different approaches. This is a brief overview with comments, acting as a foundation for further study (beyond the scope of this thesis).

This issue has been addressed by a number of authors independently. Bulmer (1991) for example suggested a confidence set based on the GLS SS from pairwise distances statistic, and its asymptotic form (the χ^2 distribution). This sort of test is of the first kind, with the repeat sampling conceptualized in multivariate normal 'distance' space. It conceptually straight forward; all trees are evaluated, and all trees with a GLS SS of less than a χ^2 variable with d.f. = {number of pairwise distances - number of fitted parameters in model are deemed within the interval while all others are outside it}. An similar approach was suggested by Navidi *et al.* (1991) in the context of linear invariants and ML. This type of confidence set will be conservative in many

real situations, when the model barely fits the data, and perhaps just one tree manages to be within the confidence set limit. A possible modification to take this conservatism into account is to (if necessary) subtract from all trees the amount required so that the best tree has the expected fit of a χ^2 random variable with the expected degrees of freedom. A useful theoretical feature of these χ^2 tests is that it is valid to build these confidence intervals (under the asymptotic assumptions of lots of data), and there is no problem in comparing the fit or likelihood of one unnested model (tree) directly with another, but rather all tree models are compared to the unconstrained model.

Unfortunately, sparseness is a real problem for the ML tests. One approach is a complete Monte Carlo simulation. As Goldman (1993a) points out these simulations can be used to compare any two models, and a different tree is a different model. However, for this use we suspect that the modification to Monte Carlo simulations suggested in section 6.2.7 (where a resampling method like the bootstrap is used to explore the range of valid Monte Carlo model parameters) is important. Otherwise the expectation is that the confidence interval of trees may become too narrow by tending to reject trees other than the ML tree which originally fitted best. This approach constitutes the second alternative mentioned in the first paragraph: comparing fit to predicted data.

An approach to estimating a confidence set of trees which does not require such strong asymptotic assumptions (although it may be biased with small samples) is to estimate a confidence set based on the results of bootstrap resampling (this is back to the approach of resampling from the original data). An early example of this approach can be found in Penny and Hendy (1986), where they estimated the expected number of trees which could be within the confidence set of the best tree. Bulmer (1991) considered a formalisation of this, whereby the data is bootstrapped trees are built, then ranked by decreasing frequency of occurrence. A confidence interval is made by summing these ranked trees until greater than $(1 - \alpha) \times 100\%$ of all trees are in the confidence set (where α is the critical value chosen).

The approach suggested by Bulmer (1991) should work well for most very small trees (say no more than six taxa) with a reasonable number of resamples (say 1000)(see Cao *et al.* 1994, for some such examples). However, for larger trees a real problem is that the number of possible trees in the confidence set can quickly grow. Consequently, unless the number of replicates also becomes very large there is a real danger of excluding many trees which should be in the confidence interval. A possible solution to this problem might be to use the bootstrap not to pick trees directly from the resampled data, but to build up a picture of the distribution of the fit statistic (e.g. likelihood) under resampling. One way would be to keep just the likelihood of the best fitting tree from each bootstrap replicate. The confidence interval for the sampled trees would then be all trees evaluated on the original sample with a likelihood at least as high as the best 95% of likelihoods in the bootstrap replicates.

Kishino and Hasegawa (1989), have suggested an approach similar to the last one. Instead of comparing all trees to an overall distribution, they ask how often does tree T_1 have a higher

likelihood than tree T_2 in bootstrap replicates? If it is not greater than in 95% of replicates then the two trees are not deemed significantly different. Taking the best tree, and making this comparison to all other trees, an approximate confidence interval could be established. Kishino and Hasegawa (1989) developed such a pairwise test which is not overly computationally expensive (and it is available in the program Phylip, Felsenstein 1993). Bin were dissatisfied with this one to one series of comparisons, Hasegawa and his associates sought other approximations to the recovery rate of different ML trees from bootstrap pseudoreplicates. Towards this end they have also derived two distinct approximations to estimating these proportions, without the need for the intensive calculations of the bootstrap (see Hasegawa and Kishino 1993 for a review). This first method is called the RELL (resampling estimated log likelihood), while the second is the MND (multivariate normal distribution) approximation method (both methods were originally derived in Kishino *et al.* 1990). The RELL method is less computationally intensive when there are few site patterns, while the MND method is less intensive if there are a few sequences very log sequences. In a simple four taxon case both approximations seemed reasonable (Hasegawa and Kishino 1993), but much more study of their behaviour is needed.

Part of the problem with the pairwise comparison approach is that it involves comparison of non-nested models (different trees), and it is open to criticism on this count. This problem makes itself most pronounced when using asymptotic results to devise a likelihood ratio test of one tree against the other. One problem is that two resolved binary trees will be equally parameterised by this criterion, making the differences in lnL apparently have zero degrees of freedom (Felsenstein 1988 suggest an ad hoc solution of allowing one degree of freedom to make the test). The bootstrap approach and the confidence set approach avoid this problem by comparing each tree directly with the data, rather than one directly against the other. Asymptotically these two tests become very similar if the model holds.

All of these approaches are easily performed in the framework of Hadamard conjugations. Optimisation of individual tress by likelihood or X^2 seems preferable for reasons of statistical efficiency, whereas tree selection by direct selection from $\hat{\gamma}$ (e.g. by parsimony) is preferable in the context of computational efficiency (although of dubious wisdom given performance in simulations which tended to be worse than direct selection form the untransformed observed data section 5.8, Charleston 1994). If Monte Carlo methods are to be used to estimate confidence sets of trees, then it seems preferable to use a modification to avoid conditioning on a specific tree and model parameters (see section 6.2.7).

6.7 TESTING WHETHER TWO DATA SETS EVOLVED BY THE SAME TREE

This next set of tests deal with the question of whether or not data sets should be combined into a single analysis. There are many ways that data can be combined: adding the data directly together: just adding tree likelihoods: consensus tree techniques: transforming the data, then adding it. All of these methods are also amenable to weighting of the different parts in an aim to reduce sampling variance, or systematic error. In all cases the first question confronting the

researcher is "is it appropriate to combine these data?" Often, it only makes sense to combine data if there is confidence that the two data sets are not disagreeing (Bull *et al.* 1993, provide an example of the troubles that can occur if incompatible data sets are added together). One measure of agreement is that bar sampling error two data could be suggesting the same thing. Here we look in detail at the types of test which can be made, starting with a general case of "did two genes evolve by the same process," then moving on to more tree specific tests.

6.7.1 Did two sets of data evolve by the same processes?

This is a very general question of whether there appears to be any reason to suggest two genes evolved differently (different mechanism or different tree). It is based upon asking "are two $\hat{\mathbf{s}}$ vectors within sampling error of each other?" The crucial assumption of this test is that sites are independent (so we have a multinomial model). There are two vectors of the observed frequency of site patterns $\hat{\mathbf{f}}(1)$ and $\hat{\mathbf{f}}(2)$ to compare, each estimate from c_1 and c_2 sites respectively (with n distinct patterns observed, so if $\hat{f}(2)_i$ and $\hat{f}(2)_i$ are both zero, this pattern i-th is not considered further). The null hypothesis is that both sets of data are a random sample from the same vector of pattern probabilities. The maximum likelihood estimate of this vector ($\mathbf{g}$) of pattern probabilities has the i-th entry estimated as $g_i = \{(\hat{f}(1)_i + \hat{f}(2)_i) / (c_1 + c_2)\}$ $((c_1 + c_2)$ (a standard result under a multinomial model, e.g. see chapter 4). To measure the departure of the two samples from the ML estimate of the expectation, it is possible to use any member of the power divergence family, (see Read and Cressie 1986) with the likelihood ratio statistic G^2 being quite suitable (this is the sum over all cells of $2(\hat{f}_i \times \ln[\hat{f}_i/(cg_i)])$). Asymptotically ($c \rightarrow \infty$) under the null hypothesis, this statistic will have a χ^2 distribution with d.f. = {2(number of distinct patterns) - 2 - (number of estimated parameters - 1)} = ($2n$ - 2 - n + 1) = n - 1. There will be no partial recovery of degrees as with ordinary ML estimators, since the estimators used here are directly based on the multinomial distribution (Stuart and Ord 1990, p. 1168).

In most real situations, these vectors will be very sparse and the asymptotic statistic cannot be relied upon. Two solutions are: run a simulation, by randomly resampling c_1 and c_2 patterns respectively from $\mathbf{g}$. If the observed G^2 statistic does not fall within the smallest $(1 - \alpha) \times 100\%$ of these resampled G^2 statistics, then the null model is rejected at an α level of 0.05. Alternatively, group cells. If cell counts are uneven, a useful rule is to group cells so that 80% have an expected value of ≥ 5, with no cells having an expected value of less than 1. A second rule, more appropriate if cell counts can be made quite even, is to group cells so that they all have an expected value of about three (to maximise the power of the test). In practice a combination is useful: firstly group all the non-bipartition patterns by the second rule, but group the often highly uneven bipartition patterns by the first rule. It is useful not to mix bipartition patterns with other patterns, since with four or more states, the bipartitions are often the most interesting.

This type of test can also be made with an X^2 Pearson statistic (i.e. (obs. - expt.)^2 /expt.), and this statistic often has more power at detecting local deviations from expected (Read and Cressie 1986). An added advantage of this statistic is that its score per cell is a good guide to how deviant that cell is individually.

6.7.2 Testing: "Did two data sets evolve according to the same weighted tree?"

Presented are a series of tests asking did "two data sets evolve by the same tree?": a critical question before combination of data. The tests described here go from the most specific (evolution on the same weighted tree) to the more general (evolution according to the same unweighted tree. In practice they would usually be used in the opposite order. The principle of these tests is the same as that of the previous test in section 6.7.1: make a joint ML estimate of the expected models parameters, and then ask if the two observed data sets appear to fit as well as random samples from this expected model.

A useful test to make is whether the optimal weighted from one set of data fits another set of data equally well? This sort of test can be used to detect departures in the evolutionary rate of one gene with respect to another. It is also an important test to make before combining sequences together to estimate a single tree, especially if edge weights are quite different, as this will cause systematic error under most models. There are a number of ways such a test could be made: we are unaware of any of these being discussed, or used in the literature (where the main interest is often comparing one tree against another). A test of two data sets evolving by the same weighted tree would be quite simple if we knew the true weighted tree of either (and hence under the null hypothesis, of both datasets). You would simply test this tree against the optimal tree of each data set and only if it fitted both data sets acceptably would the test fail to reject the null hypothesis.

Without knowing the true tree, an alternative test is to take the optimal weighted tree from one data set and compare it to the optimal weighted tree from the second data set, and visa versa. If the tree from each data is still likely from the other data set (this could be assessed with the bootstrap and a confidence set approach), we cannot reject the null hypothesis. This test however makes some assumptions which could lead to an inappropriate rejection of the null hypothesis. An implicit assumption is that either tree is an equally good representation of the true tree. This in turn assumes that the random error of both data sets is the same, which is the same as saying they have the same sequence length (assuming they evolved by the same mechanism). Additionally, since there are two tests being made, a Bonferroni correction should be made to each individual test to keep the overall significance at the required level. Additionally it makes sense to have a one sided test, so the rejection region is $\alpha/2$ for the each test of "does the optimal tree from one data set appear likely, evaluated on the other data set."

The most statistically efficient way to estimate the true tree, under this null hypothesis, is to combine the sequences and then estimate the best tree by a method such as maximum likelihood. The optimal tree from this joint data set is then compared to the best tree from both data sets, and

only if it is not significantly worse than the optimal tree of each data set (at the $\alpha/2$ level) does the test fail to reject the null hypothesis.

Earlier we had mentioned using a confidence set approach to testing one tree against the other. A more powerful approach is probably to test the optimal tree of one data set, directly against the optimal tree of the other data set. A bootstrap replication seems to be the ideal way to do this, asking the question "How often is the tree from the other data set better fitting than the optimal tree on this data set?" If this proportion is less than $\alpha/2$ then reject the other data sets tree as an equally good alternative (within sampling error). Kishino and Hasegawa's (1989) test can be substituted for the bootstrap in the interests of saving on computation.

If the data sets evolved by different mechanisms, but the same weighted tree, the test is more difficult. The crucial steps are to obtain a statistically efficient estimate of the true tree, and the mechanism of evolution for each data set. A fully statistical method such as maximum likelihood offers great flexibility here. Once the mechanism of evolution for each data set is ascertained, construct a mixed ML model. By this I mean for any weighted tree (WT) measure the overall likelihood for this tree as the likelihood of data set 1 (measured under its mechanism of evolution (m1) and WT), the likelihood of data set data set 2 (measured under its model (m2) and WT). The edge weights on WT are then iterated until the overall likelihood is maximised. Following this a search must be made across trees. Having obtained the joint optimal WT, the same sort of test described earlier is made (of course evaluating likelihoods under the appropriate mechanism for each data set). For this test to be reliable, will be dependent partly upon having accessible mechanisms to accurately described the evolutionary process in each data set.

6.7.3 Did two data sets evolve on a weighted tree with the same relative edge lengths?

This test is the same as the previous one, but allows the relaxation that the average rate of sites in one set of data is a fixed multiple of the average rate in the second data set. This can help test if neutral evolution is the dominant force in a pair of genes. This test is also useful prior to combining data (or making a joint estimate from data), especially if the weighted tree is to be used to infer relatives divergence times of genes or species. The test is most easily visulalised under a likelihood model. This model estimates a tree of relative edge weights by jointly maximising the likelihood of data set 1 measured under its mechanism of evolution (m1) and WT, plus the likelihood of data set 2 measured under its model (m2) and λWT. Here, λ is relative rate parameter (a single scalar) in the range of 0 to ∞ which should be optimised with all other parameters (and of course reoptimised on each different labeled unweighted trees. The question to ask: is WT at least $\alpha/2$ as likely as the optimal weighted tree for data set 1, and is λWT at least $\alpha/2$ as likely as the optimal tree for data set 2 alone. Only if both tests are nonsignificant do we fail to reject the null hypothesis that both data sets evolved according to the same tree and relative edge lengths, but with different average rates of substitution (we can test this by setting l to one, which gives the test in the previous section). The bootstrapping and comparing the likelihoods of the trees for each sample is one approach, while the Kishino and Hasegawa (1989) test is an alternative. Asymptotically, the variance in G^2 expected due to

random errors on edge weights is distributed as a χ^2 variable with d.f. = number of edges in the tree, making for a test simple of the difference in likelihood.

Different rates of two genes also raise the question of weighting with respect to their contribution to determining the best joint estimate. As discussed in section 4.5, the raw likelihood might give too much weight to a rapidly evolving data set which is not resolving at all well, so that most of the differences in fit between trees are due to random error. This issue needs consideration, especially if the length of the two genes is quite unequal (especially if the rapidly evolving gene is much larger), but is beyond the scope of this thesis.

6.7.4 "Did these data sets evolve on the same tree?"

This test is very important, and is a generalisation of the previous two tests. Given different genes, it is always a temptation to combine them to estimate a single tree. Even if the edge lengths are different, the data can still be added in various ways which will yield a consistent result and identify the true tree (e.g. see appendix 3.4), but if the tree is different there are no guarantees. The tests made here are again most easily posed in the context of likelihood. Rodrigo (1993) has suggested a quite different solution to this same question, and later consider what appear to be weaknesses of his test.

The first step in evaluating the probability that two data sets both evolved according to the same unweighted tree is to again make a joint and statistically efficient estimate of the true tree, under this assumption. Doing this is even simpler than the previous tests. The overall likelihood of the data under this common unweighted tree (UT) model is simply the likelihood of dataset 1 given a suitable mechanism to describe its evolution plus weighted tree TW1 (that is a specific case of TU), plus the likelihood of dataset 2 given a suitable mechanism plus weighted tree TW2 (that is again a specific case of TU). The task is then to evaluate this model for all trees and find the best fitting. After obtaining this joint estimate of the tree, the questions asked are is the optimal weighted tree on data set 1, significantly worse fitting than the best tree for this data set 2 and visa versa? If either test rejects the null hypothesis, then our overall test concludes that the two data sets did not evolve according to the same unweighted tree, and combining them would usually be unwise. Again if rates and sequence lengths in the two data sets are quite different, this raises the question of weighting their contribution to determining the joint estimate.

If the first test is rejected, it might be informative to manipulate the trees until the test is no longer rejected. This might be done by removing taxa single taxa, then pairs of taxa, then triplets of taxa, etc. to find the largest unweighted subtree that the two data sets are compatible with. This searching procedure introduces a multiple test problem, which is not easy to deal with even in more classical problems (e.g. see Miller 1990). In some instances, it is probably wiser to search for a largest compatible subtree by building it up from smaller sets of taxa. It might also be wise to try cut and paste operations to attempt to fail to reject the null hypothesis. Such assessments would be a useful step in deciding on the evidence for reticulate evolution (e.g. due to horizontal transfer) and identify the likely lineages where a swap of genetic material may have occurred.

It is also possible to envisage this sort of test applied to parsimony. Our joint estimate of the best tree under unweighted parsimony applied to the observed sequences would be the most parsimonious tree from both data sets combined together. The next step is to find the most parsimonious tree for each data set separately. Our overall most parsimonious tree must now pass two tests in order to consider it reasonable that the two data sets are consistent with indicating the same tree. Firstly the overall most parsimonious tree must have a level of bootstrap support which is at least $\alpha/2$ of that of the best tree for that data set (where α is the significance level we wish to test at). For instance if the best tree on dataset one has a bootstrap support of 40%, then the overall optimal tree must have a bootstrap support of at least $2\%/2 = 1\%$ for $\alpha = 0.05$ (and visa versa for the other data set). In place of the bootstrap, Templeton (1983) or Kishino and Hasegawa (1989) suggest tests which should behave similarly.

Ironically, the only way I can clearly think about these tests applied to parsimony is in the context of likelihood and its associated statistical concepts. David Swofford (pers. comm.) has presented a test of the "do two data sets indicate the same parsimony tree?" I am presently unsure of the relationship of my test to his test.

Having a likelihood overview is useful in a number of ways. For example, it highlights the importance of comparing support for one tree against the other, rather than just the presence of the overall optimal tree in the bootstrap confidence interval or set for each dataset. Using this latter approach, let us assume there were 50 nearly equally parsimonious trees in the confidence set. While they might contribute together 96% of all trees resampled with the bootstrap, a tree out side this set (say the tree we wish to test against the locally optimal tree) might still have a support of 80% that of the best tree. Putting it another way, the degree of resolution (or lack of it) in a data set should not be biasing the test result. The paired comparison of just two trees also reduces the sampling variance over considering just the best trees in comparing support for two or more trees.

The test of Rodrigo (1993) is somewhat different again. Two or more data sets are considered to be compatible with evolving on the same tree if the set of optimal trees from bootstrap replicates of data set 1, includes some of the same trees as are generated from the bootstrap replicates of data set 2. A real problem fro this test is that as the number of taxa grows, even well resolved trees might indicate hundreds of trees with similar bootstrap frequencies, and the number of replicates to find two identical could become huge. Secondly, there appears no easy way to put an α level on this test. The test of Rodrigo should be able to be made more generally useful by using tree comparison metrics to define both data sets encountering trees in some small region of tree space. This is not pursued here. We suspect that this test is not as powerful as the ones described earlier.

The tests described here should by useful in both diagnosing molecular evolution, and preparing data sets for phylogenetic analysis. An interesting recent finding by Cummings *et al.* (1995) is that sampling across the genome results in more reliable trees from mtDNA. This does not necessarily argue for combining different genes irrespective of their history. It does suggest

that there may be strong local systematic errors which are diluted by widespread sampling. The test of this section might detect such factors as rejection of the hypothesis that two differently located genes evolved by the same tree when in fact they did. Other tests for the combinability of data sets are being developed and illustrated, and will be described elsewhere.

6.8 CONFIDENCE LIMITS ON FEATURES OF EVOLUTIONARY MODELS

Here the question of confidence intervals on parameters in a model of evolution, is briefly considered. First to be considered is a confidence interval on a distribution shape parameter such as k (for the Γ distribution). Following this are confidence intervals on edge lengths and ratios of edge lengths, and on transition to transversion ratios.

6.8.1 Confidence limits parameters associated with the substitution mechanism

Asymptotically, it is expected that any random variable under the true model will contribute 1.d.f. to the χ^2 distribution of the fit of data to model. Additionally, as the data becomes multivariate normal, the marginal distribution of most parameters not lying on a boundary becomes univariate normal. These two factors allow a variety of tests and construction of confidence intervals. Again, there is need to be cautious of the true distribution of the χ^2 approximation with sparse data. In the context of Hadamard conjugations a similar test can be made if minimum GLS SS is used to estimate tree edge weights and select an optimal tree. Without using GLS then at low rates of change a WLS approximation may be reasonable. Also at low rates of change, after selecting an optimal tree with a linear method such as compatibility, a reasonable approximation may be to measure the G^2 fit of $c\hat{\mathbf{s}}$ to s(T). Measuring the G^2 fit of a tree to the data in this way while varying k (the shape parameter of the Γ distribution), can provide an approximate confidence intervals (say when G2 changes by more than 3.84). More secure use of this approximation will require simulation evaluations, and consideration of the effects of sparseness on this statistic. Alternatively because the difference in fit will often involve only one parameter a Kishino-Hasegawa type test on the likelihoods of the predicted s(T) vectors might be useful (and only slightly conservative).

Another non-tree statistic is the transition to transversion ratio (tr / tv). With ML optimisation and selection, this statistic can be tested in the same way as a shape parameter. After a Hadamard conjugation, this statistic can also be estimated, without needing to select a tree (see section 2.5.1). In this instance a reliable test does seem possible with moderate amounts of data. The two components of the tr / tv ratio are simply sums of bipartition elements in $\hat{\gamma}$. Given an estimate of these elements variances and covariances (see section 3.2), the variance and covariance of the Σtr(bipartitions) and the Σtv(bipartitions) can be made (as the sum of the variances and the sum of twice the covariance of all pairs of entries). Each sum is likely to be nearly normally distributed (unless rates are very low, or very high, and or sequences are short). Given multivariate normality, a useful approximation to the variance of a ratio is $\mathrm{var}[x/y] = (x/y)^2(\mathrm{var}[x]/x^2 + \mathrm{var}[y]/y^2 - 2\mathrm{cov}[x, y]/(xy))$ (Stuart and Ord 1990, p. 325). As an example, let the sum of all transition bipartitions in $\hat{\gamma}$ be 1.4 (with a standard deviation of 0.05), and the sum of

all transversion bipartitions is 0.2 (with s.d. = 0.01) and the covariance of these two sums comes to 0.00. Then the tr / tv ratio is 7.0, and the standard error of this ratio is estimated as the square root of (49 + 0.0013 + 0.0025 - 0.0007) = 0.387. An approximate 95% confidence interval on the tr / tv ratio then becomes $7 \pm 1.96 \times 0.387$ or 6.24 to 7.76. This statistic may be of some interest since it is made independent of selecting a tree, and a comparison of its performance with the estimate by ML could be interesting.

6.8.2 Confidence limits on features of weighted trees

Single edges. The basic approach to place a confidence interval (of $(1 - \alpha) \times 100\%$) on a tree edge is to assume that it has a normal distribution, estimate its standard error, and take the confidence interval to be the observed length plus or minus z standard errors (where z is chosen from a standard normal distribution such that the it occurs on the cumulative distribution at $1 - \alpha/2$). The standard error can often be estimated without need of random sampling given an estimate of the variance-covariance matrix. If we know that this edge is in the true tree, then this is a useful confidence interval, otherwise see section 4.8.2 for routines to use errors on edge length estimates to estimate the reliability of tree selection (for further details see also section 6.5.1).

A confidence interval on a sum of edges: A confidence interval of this type is useful when interested in the total amount of evolution on one tree vs another. This type of test is an extension of the question, "is the pairwise distance from this gene to its counterpart in another species greater than that for another gene, thus indicating a higher rate of substitution?" The confidence interval on a sum of edges can be made by assuming multivariate normality, and estimating the variance of this sum as the sum of the variances and twice the sum of all the possible pairs of covariances between entries in the sum. That is, if $g = \sum_{i=1}^{n} a_i x_i$, then $Var[g] = \sum a_i^2 Var[x_i] + \sum_{i \neq j} a_i a_j \operatorname{cov}(x_i, x_j)$, gives the variance of a weighted sum (Stuart and Ord, 1987, p.324)(note that if we the variance covariance matrix of just the x_i's, then this is a sum of diagonal terms plus a sum of off-diagonal elements). In the case of $\hat{\gamma}$ the question might be the sum of all entries in $\hat{\gamma}$ ($\hat{\gamma}_0$ excluded), in which case the variance will be just that of $\hat{\gamma}$. In the case of trees, but not $\hat{\gamma}$, a confidence interval on a sum of edge weights may be distorted by tree selection. This is because there tends to be a negative correlation between the fit a tree and its sum of edge lengths (this feature is used by tree selection methods such as minimum evolution, and also shows up in the case of methods such as likelihood, see section 5.3).

6.8.3 Confidence limits for a ratio of edge lengths

This type of test is useful when using trees to calibrate divergence times, or when interested in the relative rates of two or more lineages. As an example let's consider how long the edge leading to *Sulfolobus* is relative to the edge leading to *Halobacterium* when an i.r. correction of the pairwise distance data is made and the tree is reconstructed by the minimum GLS SS

criterion. The optimal tree is the archaebacteria tree (see table 5.4) with edge lengths of 0.0667 (variance = 9.41×10^{-5}) for *Sulfolobus* and 0.0813 (variance = 7.95×10^{-5}) for *Halobacterium* (while the covariance of these edges is -1.10×10^{-5}), so ratio is 0.0667 / 0.813 = 0.820. The standard error of a ratio of x / y is approximately,

$$Var[x/y] = \left\{\frac{E[x]}{E[y]}\right\}^2 \left\{\frac{Var[x]}{(E[x])^2} + \frac{Var[y]}{(E[y])^2} - \frac{2\,\mathrm{cov}[x,y]}{E[x]E[y]}\right\} \qquad (6.8.3\text{-}1)$$

(Stuart and Ord 1987, p.325). Applying this formula to our example gives the variance of x /y as 2.51×10^{-4} (and s.e. = 0.0158). As long as both *x* and *y* are approximately bivariate normal with standard errors that are relatively small compared to their mean, then the distribution of this ratio can be treated as approximately normal. An approximate 95% confidence interval for this ratio is $0.820 \pm 1.96 \times 0.158$ or 0.51 to 1.13. Thus under the assumption of the monophyly of the archaebacteria this data does not exclude the lineage leading to the "eocytes" evolving faster than that leading to the halobacteria. An alternative way to make the test of a one to one ratio is to see if the confidence interval for the difference of these two edge lengths includes zero. The difference in edge lengths is 0.0146, which has variance of $(9.41 + 7.95 - 2 \times -1.10) \times 10^{-5}$, which gives $0.0146 \pm 1.96 \times 0.0140$ and this interval does indeed include zero (notice also the slightly lower variance for the difference as opposed to the ratio).

6.8.4 Differences in p_{inv} or shape parameters from different data sets

An interesting possibly suggested in Lockhart *et al.* (1995) is that the effective proportion of invariant sites is quite different in different genes. Some ways of testing this are suggested in section 3.6.3 . If the two sets of sequences (*a* and *b*) are being compared are disjoint, and do not overlap in their phylogenetic history, then a test that two values of p_{inv} are equal simplifies to testing if the difference between two independent normal variables is zero. The appropriate confidence interval of $p_{inv}(a) - p_{inv}(b)$ is, $p_{inv}(a) - p_{inv}(b) \pm 1.96\sqrt{(\mathrm{var}[p_{inv}(a)] + \mathrm{var}[p_{inv}(b)])}$ (when using likelihood, an asymptotic estimate of one standard error from the optimal value, is the parameter value at which the lnL has dropped by 1 G^2 unit or 0.5 of a log likelihood unit). A similar test can be made to compare the shape parameter from different data sets.

6.9 COMPREHENSIVE STANDARD ERRORS FOR DIVERGENCE TIMES

If age estimates based on relative amounts of genetic evidence are to be compared with geological dates, it is appropriate that the standard errors on these times be as inclusive of all major sources of error as possible. The first publication, I am aware, of attempting to do this is Waddell and Penny (1995). For divergence times estimated from DNA sequences, four main sources of error were identified:

(1) Stochastic error,

(2) Sequencing error

(3) Error in the calibration date

(4) Ancestral polymorphism of extinct populations

(5) Fluctuations due to the tree inference method used, and the taxa selected

Stochastic error. This source of error is due to finite sample sizes leading to fluctuations in estimated edge lengths. One of the findings in Waddell *et al.* (1994) was that error on inferred edge weights tend to be quite independent at lower rates of change. Alternatively, an estimate of the covariance matrix of edge weights can be derived. Assuming independence, then if a relative divergence time is estimated as $(e_1 + e_2)/2$ / $((e_1 + e_2)/2 + e_i + .. + e_j)$ (where e_k is an edge in the tree, so this is the backbone method of section 5.3.9), this can be considered a ratio of x/y, where $x = (e_1 + e_2)/2$ and $var[x] = (var[e_1] + var[e_2])/4$, while $y = ((e_1 + e_2)/2 + e_i + .. + e_j)$ and $var[y] = (var[x] + var[e_i] + ... + var[e_j])$, with the covariance of x and y being var[x]. Variances of edges are returned by programs such as DNAML, and the overall variance of a ratio can be estimated with formula 6.8.3-1. If covariances are available and moderately reliable, then they can easily incorporated in estimating var[x] and var[y] and cov[x, y]. Probably the most important feature of this formula is that the variance of a ratio can easily become much bigger than expected from glancing at the s.d. of x and y (see Waddell and Penny 1995). *Hasegawa et al.* (1987) have used a bootstrap procedure to estimate this error, which is more expensive, but desirable in that it gives an estimate of the form of the expected distribution to the ratio of x/y (which can become markedly skewed). A nice feature of the bootstrap is that it should take into account factors such as stochastic error in the estimated distribution of rates across sites (e.g. k or p_{inv}) when ML methods with full reoptimisation are used.

Sequencing errors: This second source of error becomes especially important in estimating divergence times within species (e.g. Hasegawa *et al.* 1993, Waddell and Penny 1995). While error rates of over 1/1000 are often quoted, this need not be the reality. 16S-like rRNA sequences published in the last two years have error rates as high as 50 per 1000 sites! (5%, estimated from sequencing studies by N. Pieniazek, pers. comm.). Other systematic studies of which I am aware seem to have error rates or the order of 1/100 to 1/400 (K. Slack and K. Hurr pers. comm.). The ability to get ample amount of very clean DNA seems to be an important factor. This source of error will tend to flatten the ratio of x / y, so making divergences younger than the calibration point(s) be biased upwards, and those older will be biased downwards. Waddell and Penny (1995) suggest approximate ways of taking it into account, based on expected values. Perhaps the biggest problem with sequencing errors is that few people seem prepared to talk about them, and the confirmed range across published studies is apparently very wide (at least from 5% down to 0.02%).

Error in the calibration date: This source of error is more than just the age of the fossils, and involves guessing how much older (or possibly younger) the actual divergence time was. Waddell and Penny (1995) make comment on the important of being able to place upper and lower boundaries using biogeographic events, while Marshall (1990) considers the problem of "not having found the older fossils" from a statistical viewpoint.

Ancestral polymorphism of extinct populations: This factor probably is expected to have greatest importance for recently diverged species with long life spans and large effective population sizes (hominoids and whales might be two examples). In a population which has reached equilibrium, for loci which are not recombining, the average coalescent time for a nuclear gene is 4N times the generation time, which in many ape subspecies appears to be about 100,000 × 15 = 1.5 million years (e.g. Morrin *et al.* 1994). Making matters worse, this time has an exponential distribution (see Hudson 1990 for a review). In estimating the divergence time of humans from chimps this is adding an extra source error which is adding on to the error in x and y, and creating a probable downward bias.

If sequences are experiencing recombination, then the distribution of the coalescent time will become more narrow and closer to normal in shape (Hudson 1990). However it will not be decreased. Using inappropriate models when there is ancestral polymorphism, recombination and closely separated divergence times can bias estimated divergence times upwards (see section 5.4.5).

Fluctuations due to the tree inference method used, and the taxa selected. This set of errors is not inconsiderable, as is shown in figure 5.8. The errors shown in this figure are reflecting systematic errors amongst a closely related set of models, and the inclusion or exclusion of one lineage of taxa which (correctly or not) are suspected of being slightly less clock-like in their evolution than other species. Disturbingly, divergence times due to more precise modeling of the mechanism of evolution dropped distinctly below this range in practically all cases. Given that these are recently separated species, we need to be aware of this possibility with much older divergences (e.g. dating the origins of mammalian orders). The expectation tends to be that more adequate approximations of the mechanism of evolution will increase older divergences relative to younger ones (so depending upon where in the tree the calibration points are this will either bias upwards or downwards actual divergence times). This need not be the case for some models (especially when the distribution rates across sites are not exactly specified, or sites change their rate class). Their needs to be more study of what treatment is best to minimise the effect of fast and slow clock species. At present I feel the approach of Waddell and Penny (1995) (see also section 5.3.9) where a range is estimated by including and excluding different subsets of species is safest.

A model to incorporate all these changes: We are currently developing a mixed bootstrap / Monte Carlo method to incorporate all these factors and give an overall estimate of the fluctuation in the expected divergence time of humans and chimps estimated from mtDNA, or an amalgamated sequence of nuclear DNA. Short cuts and approximations to some of these steps are given in Waddell and Penny (1995). The most important factor, is that the standard error of estimated divergence times more than doubles from that usually calculated assuming just finite DNA sequence length (e.g. Hasegawa *et al.* 1987). Preliminary results suggest that it is not yet possible to differentiate the divergence time any more accurately than 1 million years (one standard error), although this may be refined by longer sequences, the other four factors are limiting and need refinement themselves (sequencing error does however seem to be much less

of a problem in the large hominoid studies, where Horai *et al.* 1995 for example seem to have it to a level of considerably less than 1/2000). Work with Dick Hudson is to better describe the distribution of coalescent times when recombination rates are of the same order as substitution rates.

6.10 A BAYESIAN VIEW OF PHYLOGENETIC ANALYSES

The real power of Bayesian statistics is that they offer a way of estimating the probability of a hypothesis being correct, and this approach can potentially incorporate all the relevant information in reaching this conclusion. This is indeed a very important aim, and for much of the research in empirical sciences, the whole study is aimed at refining, or sometimes refuting this probability. It is fair to say that what most biologists want out of phylogenetic studies is know how much fabric of details they can confidently build upon what is claimed to be a phylogenetic "fact". And the other side of this, is of course, that most biologists wish to avoid going too deeply into an exercise of "biological extrapolation" if the "fact" is likely to be wrong. Kishino and Hasegawa (1989) are to date one of the few studies exploring Bayesian statistics in phylogenetics.

Stuart and Ord (1990, p. 1214) give a useful overview of the different perspectives in statistics. Interestingly, they do not find Popperian philosophy and Bayesian statistics compatible. With a slight modification of what is meant by falsifyability to incorporate degrees of likelihood (see section 1.2) and a recognition of Lakatos' (1974) view of science as more a description than a prescription (in contrast to e.g. Popper 1979), I do not find any serious incompatibility of these approaches.

6.10.1 The need to integrate different sources of knowledge

Much of phylogenetics revolves around being able to resolve key issues that arise in specific context. For example are humans more closely related to chimps, than to gorillas or are chimps and gorillas together. At the other extreme of ancientness we ask questions such as are the known extreme thermophile and methanogenic archaebacteria monophyletic? In all these issues there is definitely prior information, and if you were to question a biologist familiar with the subject, priors would not be hard to find. There is also a definite prior building up based on results of analyses of different genes. The crucial question becomes how can all this information be integrated to give a reasonable assessment of the probability that an hypothesis is correct? As already discussed in section 1.2 Bayesian statistics offers one possibility.

6.10.2 Setting up the prior

The major statistical objection to Bayesian statistics appears to be the "frequentist" argument that there is no prior distribution which meets the probabilistic requirement of acting as a null distribution when there is claimed to be no prior information. A major response to this has been the concept of 'subjective priors': that is, everyone indeed has a prior which can be pinned down with the right questions. The frequentist response is usually that this type of prior is not good since it is subjective, to which the most common Bayesian reply is that subjective probabilities

are a valid and useful concept (for example see Stuart and Ord 1987, p262). The other argument on the side of the Bayesians is that irrespective of what prior you take (as long as it is not 0 or 1), then as more data is analysed, the these results rapidly determine the range of the posterior probability. It is thus often considered good practice to give a generous confidence interval on the prior, to basically satisfy any but the sternest critic of this actual number. After the analysis, the difference in the posterior given the extremes of the prior are a gauge of uncertainty.

6.10.3 A worked example based on the archaebacteria question

Let us consider how Bayesian statistics might be used in phylogenetics, and also what some of the obstacles to its use are. This discussion is coloured with the real example of whether the archaebacteria are monophyletic, versus the halobacteria tree, versus the eocyte tree. I will refer to the main protagonists in this argument also, this discussion is not meant to convey any value judgment on their research , just my impression of the usefulness or otherwise of their stance. Subjective priors also serve a very valuable role of putting people on record in the most stark way on what their expert opinion is on a subject. This can be more informative to those unfamiliar with the development of a field than any single analysis.

We shall base our priors on the non-sequence data. Backtracking to 1980, the main indications where that all archaebacteria shared certain unique biochemical features. These features included unique cross linked ether lipid membranes, and modified nucleotides in their tRNA's (e.g. see Woese and Wolfe 1985, Kandler and Zillig 1986). In 1984, Lake and colleagues claimed to be able to see features of the ribosomes of "eocytes" shared with eukaryotes. Others have claimed that the proton pumps and other photosynthetic apparatus of halobacteria share details in common with photosynthetic eubacteria. All of this data is qualitative, but my own prior is 60% in favour of archaebacteria, and 25% for eocyte and 15% for the halobacterial tree. At an unqualified guess, I imagine that a researchers such a Lake (e.g. 1986, 1988, and Rivera and Lake 1992) would disagree with my priors and might suggest 90% for the eocyte tree and 5% for the two alternatives. In contrast Karl Woese (e.g. Woese's and Fox 1977, Woese and Wolfe 1985, Woese and Olsen 1986) might well suggest 90% for archaebacteria and 5% for each alternative. I am guessing that a crude '66%' interval for the prior of informed researchers might be from about A: 80%, E: 15%, H: 5% to A: 40%, E: 40%, H: 20% (obviously being a three dimensional problem these are not very exact, and this last value being David Penny's prior), but this needs to be verified with questionnaires.

6.10.4 Integrating prior and experimental results to update hypothesis support

The next step is to evaluate the likelihoods of the three trees. In the case of two trees the relative posterior probabilities are given as,

$$\frac{Posterior\_P(T_1)}{Posterior\_P(T_2)} = \left(\frac{\Pr ior\_P(T_1)}{\Pr ior\_P(T_2)}\right) \times \left(\frac{Likelihood\_P(T_1)}{Likelihood\_P(T_2)}\right) \qquad (6.9.4\text{-}1)$$

(For an example see Hasegawa and Kishino 1989). In our case we have three alternatives and so we would write this as,

$$\frac{Posterior\_P(T_1)}{Posterior\_P(T_2)} = \frac{Prior\_P(T_1)}{Prior\_P(T_2)} \times \frac{Likelihood\_P(T_1)}{Likelihood\_P(T_2)} \qquad (6.9.4\text{-}2)$$
$$\frac{Posterior\_P(T_3)}{Posterior\_P(T_3)} \qquad \frac{Prior\_P(T_3)}{Prior\_P(T_3)} \qquad \frac{Likelihood\_P(T_3)}{Likelihood\_P(T_3)}$$

(Note, it important to use only binary trees in this type of evaluation since the hypotheses must constitute a set of mutually exclusive and exhaustive events). The final step in the analysis can be to make these relative probabilities sum to one for ease of interpretation. So in order to get the posterior probabilities of all three trees requires the relative likelihoods of these three trees. To get likelihoods, requires choosing models and in so doing we also express subjective preferences.

6.10.5 Using resampling schemes to asses the 'likelihood' of different trees

Asymptotically, under the i.i.d. assumption, and assuming the model is correct, it is possible to use the ratio of likelihoods of optimal weighted trees to estimate the likelihoods of the relative likelihoods of the two unweighted trees they represent (e.g. see Hasegawa and Kishino 1989). A simple alternative with finite data is to use either a bootstrap resampling scheme, or given confidence in the assumed mechanism of evolution a Monte Carlo resampling scheme, to decide how often one tree is preferred over another (see Hasegawa and Kishino 1989 for an example). These resampling schemes are estimating the respective 'mass' (volume times density) of the sample space in which each tree is favoured. If the method for estimating trees is not strongly biased in its selection of different trees, then any suitable tree selection procedure can be married to the bootstrap. Thus it is not necessary to use an ML method of tree reconstruction, just one which is believed to be unbiased in its favoritism for different trees (as most methods are unbiased when parallelism and convergences are expected to favour all trees about equally and / or are rare).

Choosing a model to estimate the relative likelihoods of the alternative trees under can be contentious when dealing with very ancient divergences. I don't think it is overly harsh to suggest that many of the presently implemented ML tree selection methods may be prone to give potentially strongly biased estimates of the relative likelihoods of different trees due to base composition differences, or shifting functional roles (e.g. see Lockhart 1990, Lockhart *et al.* 1992, Lockhart *et al.* 1995). An alternative is to use a method such as the invariant sites-LogDet model of chapter 3 (an ML version of mechanism with the 28 16S-like sequences would presently be computationally prohibitive, but perhaps not for long). A tree selection procedure incorporating this transformation is given in figure 3.11, with the result that the 'likelihood' of the archaebacteria tree was about 60%, while that other two trees where each about 20% (with 25% of constant sites removed). Our concern over this method (or indeed any other method applied to these very early divergences) is that as currently implemented they pay too much attention to the more rapidly evolving sites, which seem to experience the most marked shifts in base composition. Apart from the possibility of systematic biases, analyses in section 4.5 showed that the inclusion of the more rapidly evolving sites could be seriously eroding the resolving power of this method as applied to this very deep divergence.

In this instance, I expect a valid alternative is to base tree 'likelihoods' on the use of the perfect sites of table 3.8 to resolve this part of the phylogeny. There are 8 perfect characters of types one and two supporting the archaebacteria tree, but only 1 supporting the eocyte tree, and 2 supporting the halobacteria tree. There appears to be no evidence for systematic bias in this data (i.e. the two non-optimal trees both have support which is within sampling error of each other), and so a simple procedure such as maximum parsimony applied to these observed data of four 'taxa' should have little bias, or favoritism in the trees it selects. A reasonable direct estimate of the relative support for each tree can be gained from direct multinomial calculations (e.g. Felsenstein 1983b) which gives the probability of the archaebacteria tree being clearly optimal (that is, not worse or tied as) as 0.974, or at least shared optimal as 0.982 (and in a tie of n trees, each tree is recorded as being optimal 1/n times so that all outcomes still sum to one). For the halobacteria tree the support is 0.011 (or 0.017 counting ties), while for the eocyte tree the support is less than 0.001 (still 0.001 counting ties). (Note that leaving in the many uninformative sites (here 800 - 8 - 2- 1) is strictly more correct in estimating resampling support, although the difference here is small, but would slightly reduce the support estimated for the archaebacterial tree).

Thus the relative posterior likelihoods of the three trees following evaluation of the 16S-like rRNA sequences and using my priors are, archaebacterial tree 0.995, eocyte tree 0.0004, halobacterial tree 0.004. With rescaling these posterior probabilities are 0.995 for the archaebacteria tree, 0.0004 for the eocyte tree, and 0.004 for the halobacterial tree. Even with priors as extreme as 90% for the eocyte tree, which I would not support, the rescaled posterior probabilities are A: 0.966, E: 0.017, H 0.017. Clearly it would be flying in the face of reason to still favour the eocyte tree based on just this evidence.

Recently Rivera and Lake (1993) have claimed significant new evidence in favour of the eocyte tree. Their claim is based upon a shared deletion between a thermophilic methanogen and a eukaryote (an eocyte has yet to be analysed). I consider this evidence rather weak due to the very poor species sampling, and the difficulty of alignment of the ends of this deletion in different species. I do not buy into the argument that there could not have been multiple insertions or deletions in this instance. However, in the spirit of the analysis I will admit this character as favouring the eocyte tree, and give it equal weight to one of the "perfect characters" in the SSU rRNA. Assessing the likelihood of the three trees based on the now twelve character data set now gives the likelihoods as A:0.966, E: 0.017, and H: 0.017 (again evaluated by giving out support for tied optimal trees, which favours the two less supported trees). With my priors the rescaled posterior probabilities are A:0.989 , E:0.007 , and H:0.004, while even with extreme prior favoritism of the eocyte tree, the posterior probabilities are A: 0.759, E: 0.228, and H: 0.013. Again I think most people would conclude that the archaebacterial tree is clearly favoured (but not significantly). To counter any claim that the 16S-like rRNA data may be artificially favouring the archaebacteria tree due to some characters not being independent, we point out that table 3.8 also lists 3 additional with perfect "transversion characters" for the archaebacterial tree alone. All the regions used for this analysis were clearly well aligned.

Overall I feel that this type of analysis can be sensibly expanded to incorporate practically all of our present knowledge pertinent to the monophyly (or not) of the archaebacteria. Due to the extreme age of these divergences, and the knowledge that any fixed substitution must have a selective advantage in order not to undergo any back mutation in this vast period of time, I remain skeptical of the wisdom of applying detailed models to all but the most conserved regions of the molecules, and even then favour a dual parsimony type analysis with the most conserved characters (e.g. see Olsen 1988 for a nearly identical approach). I also argue that characters should also be weighted in proportion to the adequacy of their sampling; it is easy to have very conservative looking characters based on a sample of say just seven taxa, but much less likely that these same characters will be showing no signs of homoplasy when another twenty divergent taxa are assessed. It is not right to think of such a stringent approach as "throwing away data," rather we must anticipate potentially overwhelming systematic biases working with characters showing multiple changes. If a molecule such as elongation factor α can yield as few as 4 independent site changes that conform to the criteria of "perfect characters," then this is a lot of information if they favour just one tree. Given a dozen such genes, with diverse taxa sampling, it seems likely there would be little doubt left.

A recent complication is that the four main lineages may not conform to a tree model, but have evolved with major reticulations. This is a reasonable hypothesis which may have some supporting evidence (e.g. Golding and Gupta 1995). However, there is also a good chance that much of this evidence could be the result of comparisons of parlous genes, gene migrations to the nucleus of eukaryotes which postdate the establishment of mitochondria, or just biases misleading simple models applied to a very heterogeneous mixture of sites. Even if there are reticulate events making for more than one possible resolution of the question are the archaebacteria monophyletic, we must still ask where specific major components appear to have come from and what the support for this new hypothesis is.

6.11 DISCUSSION

This is an overview of some of the tests being developed. Some have been illustrated (e.g. in Waddell and Penny 1995) while others are being worked on. It is important to develop robust statistical quickly in the rapidly expanding field of phylogenetics. For example, a recent claim in Lockhart *et al.* (1995) that some anciently diverged genes have very different proportions of invariant sites and this may be distorting results of phylogenetic analyses applied tests based on an invariant sites model. It will be important to check this test (see section 3.6.3) as it could be biased due to the true distribution of rates across sites not being invariant versus i.i.d. but perhaps inverse Gaussian distributed. If this were the case there would be a bias (which would not go away with increasing sample size) if the divergence of sequences, the trees, etc. in the two groups of sequences being compared were different. This is important to check, but seems unlikely to overturn this particular result since the difference is quite large. Rule of thumb statistics, while sometimes biased, are often the most useful in practice.

The development of Bayesian approaches is also very important as provide a framework that goes beyond a single analysis (or paper) and draw researchers into seriously considering the true extent of their knowledge. My suspicion is that if a pole of priors shows a strongly multimodal distribution, this is a good indication that the area of study is particularly dynamic and may be suffering form a lack of an agreed methodology, data, or perhaps some other inadequacy. This should be a clear warning to other scientists not to place to much confidence in strong claims either way. Hopefully, a development of Bayesian views of the field of phylogenetics might help to reduce the problem of when is a test a test, given the multiple tests problem that faces every analysis (e.g. see Rodrigo *et al.* 1994). The most positive thing is that the field is a rapidly developing one, so statistical tests should not be starved of adequate data for long, which is a great impetus for their development.

CHAPTER 7:

DISCUSSION AND OVERVIEW

7.1 INTRODUCTION

This chapter looks especially at some of the leads this thesis has opened up and which should provide interesting results in the near future. A number of themes become apparent. A practical theme is the continued development and evaluation of ML methods in phylogenetics which is driven along by both advances in computational hardware and a renewed interest in the property of statistical efficiency. Further, with computational power available, often the most straightforward way to study a new method is in a likelihood framework. This usually involves predicting the data under the model, and so allows systematic studies of a methods properties (e.g. Reeves 1992, Yang *et al.*, 1993, section 5.3). This interest in likelihood is also 'good' in that it forces us to look at some of the more fundamental issues in phylogenetics, such as precisely how the data are being interpreted.

One whole issue which is not well developed in phylogenetics is the concept of 'information.' That is, given the complex nature of the evolutionary process, at which different levels of simplification (e.g. the i.i.d. assumption, or reduction to just pairwise distances) do we discard detail of importance to describing particular features of the evolutionary process. Penny (1982) addressed this question in terms of the number of distinct quantities present in i.i.d. sequences versus i.i.d. distances (he did not consider evaluating also the covariance matrix of the distances, which contains information additional to the distances themselves). His main question still remains largely unanswered: what information relevant to phylogenetics do we loose doing this?

Another issue is what sort of information is lost in ignoring site correlations, and how does this affect i.i.d. methods. This question has recently prompted development of models taking account of, at least local correlations between nucleotide sites (e.g. Adachi and Hasegawa 1991, Felsenstein 1994, Gaut and Lewis 1994, Schöniger and von Haeseler 1994, Yang and Goldman 1994). One of the nice things about ML (and other methods which are required to predict the expected data, like minimum X^2 methods) are that they make you aware of the 'sufficient' statistic for the data, that is the minimal data form which still contains all the information relevant to the model.

Thus the concept of information content, or more appropriately 'extractable' or 'detectable' information, may well underpin reaching the type of global understanding of phylogenetic methods which is so much lacking at the moment. A crucial first step in developing a greater understanding of the 'information' in molecules is better diagnosis of their properties, a area which chapter 3 concentrated upon (and which is both distinct from, and complimentary to,

studies of structure and function, e.g. Gutell *et al.* 1985). Asking the question of what do I want from this data, and imagining what you would like to know from the data, are important steps to realising the potential information in sequences accessible through more inclusive models.

7.2 QUESTIONS FOR THE FUTURE

This section concludes the thesis by briefly looking at future directions which this work has helped open up.

Chapter 1 began by giving an overview of the relationships of phylogenetic methods. This thesis has delved into some of their relationships, but clearly there are many further questions to be answered. A nagging question which remains is when do distance methods and ML outperform each other, and when are they outperformed by parsimony or compatibility on observed sequences. The answer to this will probably have a lot to do with when estimates of the variances and covariances of edge length inferences (e.g. in $\hat{\gamma}$) become generally useful, and do not overly suffer from sampling error themselves. An important related issue is how much can the performance of each of these methods be enhanced in realistic situations by supplemental weighting or data editing schemes, and this includes the performance of likelihood methods.

Chapter 2, introduced the use of Hadamard conjugations modeling unequal rates across sites. It will be interesting to see how fine the detail recoverable by using mixed distributions will be. Alternatively it may turn out that without editing data into sites, the most useful approach is to use just a few standard distributions like the Γ and the inverse Gaussian mixed with invariant sites. This chapter also introduced order 2t-1 4-state Hadamard conjugations. Their use has just begun and what they gain and / or loose from condensing patterns down to bipartitions, is still uncertain. It may be possible to extend their use to amino acid sequences, this requires more study.

Chapter 3, shows the LogDet to be a particularly useful distance estimator, and so its analogue, the general i.r. / i.i.d. ML method of Barry and Hartigan (1987a) should also yield useful results. Vexing questions include: how long will sequences need to be before it outperforms LogDet? To what extent does this ML method with essentially free form transition matrices on each edge gain robustness to factors such as unequal rates over sites? This chapter also raised the question of how well we are using the information in ancient molecules. The most conservative "informative" changes deserve further study, as do factors such as transitions apparently being more common amongst the most conserved parsimony informative sites. It is distressing when the most conservatives sites suggest one tree (e.g. the archaebacterial tree) but subsets of all sites combined with i.i.d. methods can strongly support other trees (e.g. the eocyte tree, see Olsen and Woese 1989, section 3.6, section 5.3).

Questions raised by chapter 4 include: what sort of information might the Hadamard conjugation be best at extracting when the data did not evolve in a tree like manner? The other side of that coin is: what crucial information do distances loose when the data does not fit an assumed mechanism, and so what do signals in the distance Hadamard really mean? That is,

does a signal grouping a set of species have any direct correspondence to their being an excess of site changes supporting that grouping, or might it just be an artifact? Section 4.5 raised the important question of the usable information content of sequences when rates across sites vary. Finding the set of sites most informative to a question and excluding sites which are unlikely to be reliable are critical questions to extracting the most useful information from sequences. A related question is when (and how) to treat protein coding DNA sequences as independent sites, when to pool the coding positions into amino acids, and when to use the 61 non-stop codons as states. It would appear that at short distances, treating the codon positions as independent might yield the best signal-to-noise ratio. As amounts of substitution increase, then at some the reduction in parallelisms and convergences obtained by using more states probably becomes most important. The 61-state model should contain the information of both 4 and 20-state models, but using it in an ML or distance context involves inferring many more parameters and so increasing the variance of estimates.

From chapter 5, some questions are: how well will tree selection from $\hat{\gamma}$ be improved by weighting with the inverse of estimated variances? Will tree selection from $\hat{\gamma}$ then compete with other methods, or might it still suffer form the large variances of the longest pathsets? Getting branch and bound for i.i.d. ML further developed seems an important step, and how to most effectively take into account sequence sites not already in a subtree is a crucial question. ML with unequal rates across sites raises the possibility of more effective diagnosis of overall distributions of site rates, but it will be important to compare the fit of different distributions (e.g. Γ, inverse Gaussian, these mixed with invariant sites) before concluding any one is most common or useful. ML models of covarion type evolution are very interesting, and how to get the likelihood models in section 5.3 to take into account more continuous time switches of rate (not just at internal nodes) will be interesting. Section 5.4 showed there is useful information to be gained by using ML methods to infer reticulate phylogenies. Confirming the proposed "trapped" ancestral polymorphism in nuclear sequences will be crucial to understanding the population histories at the time prior to the emergence of hominids and ultimately humans.

The inconsistency of i.r. ML in the Felsenstein zone was not unexpected, but its occurrence under a molecular clock was more surprising. It will be interesting to see if there are many clock-like situations where ML is inconsistent. Another way of looking at this question is will ML still be markedly more robust than other commonly used criteria under a clock. The suggestion so far appears to be yes it will. Confirming the importance of the anti-Felsenstein zone in real studies will be critical to its acceptance as a real problem. It is possible that covarion models may cause this sort of 'overcorrection' with i.i.d. methods. This needs more study.

Maximising the efficiency (both computational and statistical) and power of statistical tests is a major challenge. It will be interesting to see how the tests described in chapter 6 perform by these criteria. Hopefully, resolving edges in trees will also take on a more Bayesian flavour, and the use of quantified prior information is important. Perhaps the signal from the most conserved

informative sites might play the role of evidence which all parties will accept, as it may be least dependent upon modeling the uncertain effects of multiple hits.

Overall then, the thesis has helped to answer some questions, and has raised just as many. The field of phylogenetics is a buzzing at the moment, and the anticipation must be some very interesting developments in statistical sequence analysis in the next few years.

Bibliography

ADACHI, J., AND M. HASEGAWA. (1992). Computer Science Monographs, No. 27. MOLPHY: Programs for molecular phylogenetics, I.-PROTML: Maximum likelihood inference of protein phylogeny. Institute of Statistical Mathematics, Tokyo.

ADACHI, J., AND M. HASEGAWA. (1995). Improved dating of the human-chimpanzee separation in the mitochondrial DNA tree: Heterogeneity among amino acid sites. J. Mol. Evol. (in press).

AGRESTI, A. (1990). Categorical data analysis. John Wiley and sons, New York.

AKAIKE, H. (1973). Information theory and an extension of the maximum likelihood principal. *In* 2nd International Symposium of Information Theory (ed. B. N. Petrov, and F. Csaki). Akademiai Kiado, Budapest.

ARCHIE, J.W. (1989a). A randomization test for phylogenetic information in systematic data. Syst. Zool. 38: 239-252.

ARCHIE, J.W. (1989b). Phylogenies of plant families: A demonstration of phylogenetic randomness in DNA sequence data derived from proteins. Evolution 43: 1796-1800.

ARCHIE, J.W., C. SIMON AND A. MARTIN. (1989). Small sample size does decrease the stability of dendrograms calculated from allozyme-frequency data. Evolution 43: 678-683.

BAILEY, W. J., K. HAYASAKA, C. G. SKINNER, S. KEHOE, L. C. SIEU, J. L. SLIGHTOM, AND M. GOODMAN. (1992). Reexamination of the African Hominoid trichotomy with additional sequences from the primate β-globin gene cluster. Mol. Phyl. Evol. 1:97-135.

BAKER, M.D., C.R. VOSSBRINCK, J.V. MADDOX, AND A.H. UNDEEN. (1994). Phylogenetic relationships among *Vairimorpha* and *Nosema* species (Microspora) based on ribosomal RNA sequence. J. Invert. Pathol. 64: 100-106.

BALDAUF, S.L., AND J.D. PALMER. (1993). Animals and fungi are each others closest relatives: Congruent evidence from multiple proteins. Proc. Natl. Acad. Sci. USA 90: 11558-11562.

BANDELT, H.-J., AND A.W.M. DRESS. (1992). Split decomposition: A new and useful approach to phylogenetic analysis of distance data. Mol. Phyl. Evol. 1: 242-252.

BAR-HEN, A., AND D. PENNY. (1996). Estimating the bias on the LogDeterminant transformation for evolutionary trees. Applied Mathematics Letters (in press).

BARRY, D., AND J.A. HARTIGAN. (1987a). Statistical analysis of Hominoid molecular evolution. Statistical Science. 2: 191-210.

BARRY, D., AND J.A. HARTIGAN. (1987b). Asynchronous distance between homologous DNA sequences. Biometrics 43: 261-276.

BELLMAN, R. (1970). Introduction to matrix analysis. McGraw-Hill, New York.

BLAISDELL, B.E. (1985). A method for estimating from two aligned present-day sequences, their ancestral composition and subsequent rates of substitution, possibly different in the two lineages, corrected for multiple and parallel substitutions at the same site. J. Mol. Evol. 22: 69-81.

BOORSTEIN, W.R., T. ZEIGELHOFFER, AND E.A. CRAIG. (1994). Molecular evolution of the HSP70 multigene family. J. Mol. Evol. 38: 1-17.

BOWCOCK, A.M., J.R. KIDD, J.L. MOUNTAIN, J.M. HEBERT, L. CAROTENUTO, K.K. KIDD, AND L.L. CAVALLI-SFORZA (1991). Drift, admixture, and selection in human evolution: A study with DNA polymorphisms. Proc. Natl. Acad. Sci. USA 88: 839-843.

BOX, G.E.P., AND D.R. COX. (1964). An analysis of transformations. Journal of the Royal Statistical Society B, 26: 211-252.

BROWN, J.R., AND W.F. DOOLITTLE. (1995). Root of the universal tree of life based on ancient aminoacyl-tRNA synthase gene duplications. Proc. Natl. Acad. Sci. USA 92: 2441-2445.

BRUL, S., AND C.K. STUMM. (1994). Symbionts and organelles in anaerobic protozoa and fungi. Trends Ecol. Evol. 9: 319-324.

BRUNS, T.D., R. VILGALYS, S.M. BARNS, D. GONZALEZ, D.S. HIBBETT, D.S. LANE, L. SIMON, S. STICKEL, T.M. SZARO, W.G. WEISBURG, AND M.L. SOGIN. (1992). Evolutionary relationships within the fungi: Analyses of nuclear small subunit rRNA sequences. Mol. Phyl. Evol. 1: 231-241.

BULL, J. J., J. P. HUELSENBECK, C. W. CUNNINGHAM, D. L. SWOFFORD AND P. J. WADDELL. (1993). Partitioning and combining data in phylogenetic analysis. Syst. Biol. 42: 384-397.

BULMER, M. (1991a). Use of the method of generalised least squares in reconstructing phylogenies from sequence data. Mol. Biol. Evol. 8:868-883.

BULMER, M. (1991b). Strand symmetry of mutation rates in the β-globin region.. J. Mol. Evol. 33: 305-310.

BUNEMAN, P. (1971). The recovery of trees from measures of dissimilarity. *In* Mathematics in the Archaeological and Historical Sciences (ed. F. R. Hodson, D. G. Kendall, and P. Tautu), p. 387-395. Edinburgh University Press.

BURGGRAF, S., A. CHING, K.O. STETTER, AND C.R. WOESE. (1991). The sequence of *Methanospirillum hungatei* 23S rRNA confirms the specific relationship between the extreme halophiles and the methanomicrobiales. System. Appl. Microbiol. 14: 358-363.

CACCONE, A., AND J. R. POWELL. (1989). DNA divergence among hominoids. Evolution 43: 925-942.

CANN, R. L., M. STONEKING AND A.C. WILSON. (1987). Mitochondrial DNA and human evolution. Nature 325: 31-36.

CANNING, E.U. (1994). Sexual processes of Microsporidia. Abstract to the 1994 "International Conference on Protistology," held at Halifax, Nova Scotia.

CAVALIER-SMITH, T. (1993). Kingdom Protozoa and its 18 phyla. Microbiological Reviews 57: 953-994.

CAVALLI-SFORZA, L. L. AND A. W. F. EDWARDS. (1967). Phylogenetic analysis: Models and estimation procedures. Evolution 32:550–570 and Am. J. Hum. Genet. 19:233-257.

CAVENDER, J. A. 1978. Taxonomy with confidence. Math. Biosci. 40: 271-280 (erratum, 44: 309, 1979).

CAVENDER, J.A., AND J. FELSENSTEIN. (1987). Invariants of phylogenies in a simple case with discrete states. Journal of Classification 4: 57-71.

CHAKRAVARTI, A., K.H. BUETOW, S.E. ANTONARAKIS, P.G. WABER, C.D. BOEHM, AND H.H. KAZAZIAN. (1984). Nonuniform recombination within the human β–globin gene cluster. Am. J. Hum. Genet. 36: 1239-1258.

CHANG, J.T., AND J.A. HARTIGAN. (1991). Reconstruction of evolutionary trees from pairwise distributions on current species. Interface 254-257 (publication does not have volume numbers).

CHARLESTON, M.A. (1994). Factors affecting the performance of phylogenetic methods. Ph. D. Thesis. Massey University.

CHARLESTON, M.A., M.D. HENDY, AND D. PENNY. (1993). Neighbor-joining uses the optimal weight for net divergence. Mol. Phyl. Evol. 2: 6-12.

CHARLESTON, M.A., M.D. HENDY, AND D. PENNY. (1994). Effects of sequence length, tree topology, and number of taxa on the performance of phylogenetic methods. J. Comp. Biol. 1: 133-151.

CHIPPINDALE, P. T., AND J.J. WIENS. (1994). Weighting, partitioning, and combining characters in phylogenetic analysis. Syst. Biol. 43: 278-277.

CHURCHILL, G. A., VON HAESELER, A., AND W. C. NAVIDI. (1992). Sample size for phylogenetic inference. Mol. Biol. Evol. 9: 753-769.

COCHRAN, W.G. (1963). Sampling techniques, second edition. John Wiley and Sons, New York.

CUMMINGS, M.P., S.P. OHTA, AND J. WAKELEY. (1995). Sampling properties of DNA sequence data in phylogenetic analysis. Mol. Biol. Evol. 12: 814-822.

DAMS, E., L. HENDRIKS, Y. VAN DE PEER, J. M. NEEFS, G. SMITS, I. VANDENBEMT, AND R. DE WACHTER. (1988). Compilation of small subunit RNA sequences. Nucleic Acids Research 16[Suppl.]: r87-r174.

DEBRY, R.W. (1992). The consistency of several phylogeny-inference methods under varying evolutionary rates. Mol. Biol. Evol. 9: 537-551.

DEBRY, R.W., AND L.G. ABLE (1995). The relationship between parsimony and maximum-likelihood analyses: Tree scores and confidence estimates for three real data sets. Mol. Biol. Evol. 12: 291-297.

DRESS, A., M. D. HENDY, D. HUSON, P. J. LOCKHART, D. PENNY, M. A. STEEL, AND P. J. WADDELL. (1995). The validity of correcting distances from sequence data for tree building. (in preparation).

DUECK, G. (1990). New Optimisation Heuristics: The Great Deluge algorithm and Record-to-Record travel. Scientific Center technical report, IBM Germany, Tiergartenstrasse 15, D-6900 Heidelberg, Germany.

EDWARDS, A.W. F. (1972). Likelihood. Cambridge University Press, Cambridge.

EDWARDS, A.W.F. (1992). Likelihood, expanded edition. Johns Hopkins University Press, Baltimore.

EERNISSE, D.J., AND A.G. KLUGE. (1993). Taxonomic congruence versus total evidence, and the phylogeny of amniotes inferred from fossils, molecules and morphology. Mol. Biol. Evol. 10:

EFRON, B. (1982). The jackknife, the bootstrap, and other resampling plans. CBMS-NSF Regional Conference Series in Applied Mathematics, Monograph 38. Soc. Indust. Appl. Math., Philadelphia.

EFRON, B. AND G. GONG. (1983). A leisurely look at the bootstrap, the jackknife, and cross-validation. Am. Statist. 37:36-48.

EFRON, B., AND R. TIBSHIRANI. (1986). Bootstrap methods for standard errors, confidence intervals, and other measures of statistical accuracy. Statistical Science 1: 54-77.

EVANS, S.N., AND T.P. SPEED. (1993). Invariants of some probability models used in phylogenetic inference. Annals of Statistics 21: 355-377.

FARMER, M.A. (1993). Ultrastructure of *Ditrichomonas honigbergii* N. G., N. Sp. (Parabasalia) and its relationship to amitochondrial protists. J. Euk. Microbiol. 40: 619-626.

FARRIS, J.S. (1969). A successive approximations approach to character weighting. Syst. Zool. 16: 44-51.

FARRIS, J.S. (1973). A probability model for inferring evolutionary trees. Syst. Zool. 22: 250-256.

FARRIS, J.S. (1981). Distance data in phylogenetic analysis. *In* Advances in Cladistics: Proceedings of the First Meeting of the Willi Hennig Society (ed. V. A. Funk and D. R. Brooks), p. 3-23. New York Botanical Garden, Bronx.

FARRIS, J.S. (1985). Distance data revisited. Cladistics 1: 67-85.

FARRIS, J.S. (1986). Distances and cladistics. Cladistics 2: 144-157.

FELSENSTEIN, J. (1973). Maximum likelihood and minimum-steps methods for estimating evolutionary trees from data on discrete characters. Syst. Zool. 22: 240-249.

FELSENSTEIN, J. (1978a). Cases in which parsimony or compatibility methods will be positively misleading. Syst. Zool. 27: 401-410.

FELSENSTEIN, J. (1978b). The number of evolutionary trees. Systematic Zoology 27: 27-33.

FELSENSTEIN, J. (1981a). Evolutionary trees from DNA sequences: a maximum likelihood approach. J. Mol. Evol. 17:368-376.

FELSENSTEIN, J. (1981b). A likelihood approach to character weighting and what it tells us about parsimony and compatibility. Biological journal of the Linnnean society 16: 183-186.

FELSENSTEIN, J. (1982). Numerical methods for inferring evolutionary trees. Quart. Rev. Biol. 57: 379-404.

FELSENSTEIN, J. (1983a). Inferring evolutionary trees from DNA sequences. In: Statistical analysis of DNA sequence data (ed. B. S. Weir). Marcel Dekker, Inc. New York.

FELSENSTEIN, J. (1983b). Confidence limits on phylogenies with a molecular clock. Syst. Zool. 34: 152-161.

FELSENSTEIN, J. (1984). Distance methods for inferring phylogenies: A justification. Evolution 38: 16-24.

FELSENSTEIN, J. (1985a). Confidence limits on phylogenies: An approach using the bootstrap. Evolution 39:783-791.

FELSENSTEIN, J. (1985b). Confidence limits on phylogenies with a molecular clock. Syst. Zool. 34: 152-161.

FELSENSTEIN, J. (1986). Distance methods: A reply to Farris. Cladistics 2: 130-143.

FELSENSTEIN, J. (1987). Estimation of hominoid phylogeny from a DNA hybridization data set. J. of Mol. Evol. 26: 123-131.

FELSENSTEIN, J. (1988). Phylogenies from molecular sequences: Inference and reliability. Ann. Rev. Genet. 22: 521-565.

FELSENSTEIN, J. (1992a). Estimating effective population size from samples of sequences: inefficiency of pairwise and segregating sites as compared to phylogenetic estimates. Genet. Res. Camb. 59: 139-147.

FELSENSTEIN, J. (1993). *PHYLIP (Phylogeny Inference Package) and manual*, version 3.5c. Department of Genetics, University of Washington, Seattle.

FELSENSTEIN, J. AND H. KISHINO. (1993). Is there something wrong with the bootstrap on phylogenies? A reply to Hillis and Bull. Syst. Biol. 42: 193-200.

FITCH, W.M., AND E. MARGOLIASH. (1967). Construction of phylogenetic trees. Science 155: 279-284.

FITCH, W.M., AND E. MARKOWITZ. (1970). An improved method for determining codon variability in a gene and its application to the rate of fixation of mutations in evolution. Biochemical Genetics 4: 579-593.

FLEMING, T.R. (1982). One-sample multiple testing procedure for phase II clinical trials. Biometrics 38: 143-151.

FREUND, J.E., and R.E. WALPOLE. 1987. Mathematical statistics. 4th edition. Prentice-Hall International, London.

FU, Y.X., AND M.A. STEEL (1995). Classifying and counting linear phylogenetic invariants for the Jukes-Cantor model. J. Comp. Biol. 2: 39-47.

FUKAMI-KOBAYASHI, K. AND Y. TATENO. (1991). Robustness of maximum likelihood tree estimation against different patterns of base substitutions. J. Mol. Evol. 32: 79-91.

GARRET, R.A., C. AAGAARD, M. ANDERSEN, J.Z. DALGAARD, J. LYKKE-ANDERSEN, H.T.N. PHAN, S. TREVISANATO, L. ØSTERGAARD, N. LARSEN, AND H. LEFFERS. (1994). Archael rRNA operons, intron splicing and homing endonucleases, RNA polymerase operons and phylogeny. System. Appl. Microbiol. 16: 680-691.

GOGARTEN, J. P., H. KIBAK, P. DITTRICH, L. TAIZ, E. J. BOWMAN, B. J. BOWMAN, M. F. MANOLSON, R. J. POOLE, T. DATE, T. OSHIMA, J. KONISHI, K. DENDA, AND M. YOSHIDA. (1989). Evolution of the vacuolar H^+-ATPase: Implications for the origin of eukaryotes. Proc. Natl. Acad. Sci. 86: 6661-6665.

GOJOBORI, T., E. N. MORIYAMA AND M. KIMURA. (1990). Statistical methods for estimating sequence divergence. Methods in Enzymology. 183: 531-550.

GOLDING, G.B. (1983). Estimates of DNA and protein sequence divergence: An examination of some assumptions. Mol. Biol. Evol. 1: 125-142.

GOLDING, G.B., AND R.S. GUPTA (1995). Protein based phylogenies support a chimeric origin for the eukaryotic genome. Mol. Biol. and Evol. 12: 1-6.

GOLDMAN, N. (1990). Maximum likelihood of phylogenetic trees, with special reference to Poisson process models of DNA substitution and to parsimony analysis. Syst. Zool. 39: 345-361.

GOLDMAN, N. (1993a). Statistical tests of models of DNA substitution. J. Mol. Evol. 36: 182-198.

GOLDMAN, N. (1993b). Simple diagnostic tests of models of DNA substitution. J. Mol. Evol. 37: 650-661.

GOLDMAN, N. AND Z. YANG. (1994). A codon-based model of nucleotide substitution for protein-coding DNA sequences. Mol. Biol. Evol. 11: 725-736

GOODALL, J. (1990). Through a window: Thirty years with the chimpanzees of Gombe. Weidenfeld and Nicolson, London.

GOUY, M., AND W.-H. LI. (1988). Nature 331: 184-186

GOUY, M. AND W.-H. LI. (1989a). Phylogenetic analysis based on rRNA sequences supports the archaebacterial rather than the eocyte tree. Nature 339: 145-147.

GOUY, M. AND W.-H. LI. (1989b). Molecular phylogeny of the kingdoms Animalia, Plantea, and Fungi. Mol. Biol. Evol. 6: 109-122.

GOUY, M., AND W.-H. LI. (1990). Gouy and Li reply (to Lake). Nature 343: 419.

GRAHAM, R.C., AND L.R. FOULDS. (1982). Unlikelihood that minimal phylogenies for a realistic biological study can be constructed in reasonable computational time. Mathematical Biosciences 60: 133-142.

GUTELL, R.R., B. WEISER, C.R. WOESE, AND H.F. NOLLER. (1985). Comparative anatomy of 16S-like ribosomal RNA. Prog. Nucleic. Acid Res. Mol. Biol. 32: 155-216.

HALL, P. AND M.A. MARTIN. (1988). On bootstrap resampling and iteration. Biometrika 75: 661-671.

HAN, T-M., AND B. RUNNEGAR. (1992). Megascopic eukaryotic algae from the 2.1-billion-year-old Negaunee iron-formation, Michigan. Science 257: 232-235.

HASEGAWA, M., A. DI RIENZO, T.D. KOCHER, AND A.C. WILSON. (1993). Toward a more accurate time scale for the human mitochondrial DNA tree. J. Mol. Evol. 37: 347-354.

HASEGAWA, M., AND M. FUJIWARA. (1993). Relative efficiencies of the maximum likelihood, maximum parsimony, and neighbor-joining methods for estimating protein phylogeny. Mol. Phyl. Evol. 2: 1-5.

HASEGAWA, M., AND T. HASHIMOTO (1993). Ribosomal RNA trees misleading? Nature 361: 23

HASEGAWA, M., T. HASHIMOTO, AND J. ADACHI. (1992). Origin and evolution of eukaryotes as inferred from protein sequence data. *In*: The origin and evolution of the cell (ed. H. Hartman, and K. Matsuno). World Scientific Publishing Co., Singapore.

HASEGAWA, M., T. HASHIMOTO, J. ADACHI, N. IWABE, AND T. MIYATA (1993). Early branchings in the evolution of eukaryotes: Ancient divergence of *Entamoeba* that lacks mitochondria revealed by protein sequence data. J. Mol. Evol. 36: 380-388.

HASEGAWA, M., AND H. KISHINO. (1989). Confidence limits on the maximum-likelihood estimate of the hominoid tree from mitochondrial-DNA sequences. Evolution. 43: 672-677.

HASEGAWA, M., H. KISHINO AND N. SAITOU. (1991). On the maximum likelihood method in molecular phylogenetics. J. Mol. Evol. 32: 443-445.

HASEGAWA, M., H. KISHINO, AND T. YANO. (1985). Dating of the human-ape splitting by a molecular clock of mitochondrial DNA. J. Mol. Evol. 21: 160-174.

HASEGAWA, M., H. KISHINO, AND T. YANO. (1989). Estimation of branching dates among primates by molecular clocks of nuclear DNA which slowed down in Hominoidea. J. Mol. Evol. 18: 461-476.

HASHIMOTO, T., J. ADACHI, AND M. HASEGAWA. (1994a). Phylogenetic place of *Giardia lamblia*, a protozoan that lacks mitochondria. Endocytobiosis and cell research (in press).

HASHIMOTO, T., Y. NAKAMURA, F. NAKAMURA, T. SHIRAKURA, J. ADACHI, N. GOTO, K. OKAMOTO, AND M. HASEGAWA. (1994b). Protein phylogeny gives a robust estimation for early divergences of eukaryotes: Phylogenetic place of a mitochondrion-lacking protozoan *Giardia Lamblia*. Mol. Biol. Evol. 11: 65-71.

HEIN, J. (1990). A unified approach to alignments and phylogeny reconstruction. ed. R. Doolittle. Methods in Enzymology. 183: 626-645.

HENDY, M.D. (1989). The relationship between simple evolutionary tree models and observable sequence data. Syst. Zool. 38: 310-321.

HENDY, M.D. (1991). A combinatorial description of the closest tree algorithm for finding evolutionary trees. Discrete Mathematics 96: 51-58.

HENDY, M.D., AND M.A. CHARLESTON. (1993). Hadamard conjugation: A versatile tool for modeling nucleotide sequence evolution. New Zealand Journal of Botany (Conference Issue) 31: 231-238.

HENDY, M.D., and D. PENNY. (1982). Branch and bound algorithms to determine minimal evolutionary trees. Mathematical Biosciences 59: 277-290.

HENDY, M.D., and D. PENNY. (1989). A framework for the quantitative study of evolutionary trees. Syst. Zool. 38: 297-309.

HENDY, M.D., and D. PENNY. (1993). Spectral analysis of phylogenetic data. Journal of Classification 10: 5-24.

HENDY, M.D., and D. PENNY. (1996). Complete families of linear invariants for some stochastic models of sequence evolution, with and without the molecular clock assumption. J. Comp. Biol. (in press).

HENDY, M.D., D. PENNY, AND M.A. STEEL. (1992). Discrete Fourier analysis for evolutionary trees. Mathematical and Information Sciences Report, Series B: 92/2, Massey University, Palmerston Nth, New Zealand (ISSN 1171-7637).

Appears also as article below with 2 illustrative examples (one referred to in this thesis) removed due to space restrictions.

HENDY, M.D., D. PENNY, AND M.A. STEEL. (1994). Discrete Fourier analysis for evolutionary trees. Proc. Natl. Acad. Sci. USA. 91: 3339-3343.

HENNIG, W. (1966). Phylogenetic Systematics (translated by D.D. Davis and R. Znagerl). University of Illinois Press, Urbana.

HILLIS, D.M. (1991). Discriminating between phylogenetic signal and random noise in DNA sequences. *In* Phylogenetic Analysis of DNA Sequences, p. 278-294 (ed. M. M. Miyamoto and J. Cracraft). Oxford University Press, New York.

HILLIS, D.M., AND J.J. BULL. (1993). An empirical test of bootstrapping as a method for assessing confidence in phylogenetic analysis. Syst. Biol. 42: 182-192.

HILLIS, D., J.P. HUELSENBECK, AND C.W. CUNNINGHAM. (1994). Application and accuracy of molecular phylogenies. Science 264: 671-677.

HILLIS, D.M., J.P. HUELSENBECK, AND D.L. SWOFFORD. (1994). Hobgoblin of phylogenetics? Nature 369: 363–364.

HINKLE, G., AND M.L. SOGIN. (1993). The evolution of the Vahlkampfiidae as deduced from 16S-like ribosomal RNA analysis. J. Euk. Microbiol. 40: 599-603.

HORAI, S., K. HAYASAKA, R. KONDO, K. TSUGANE, AND N. TAKAHATA. (1995). Recent African origin of modern humans revealed by complete sequences of hominoid mitochondrial DNAs. Proceedings of the National Academy of Sciences (USA). 92: 532-536.

HORAI, S., Y. SATTA, K. HAYASAKA, R. KONDO, T. INOUE, T. ISHIDA, S. HAYASHI AND N. TAKAHATA. (1992). Man's place in the Hominoidea revealed by mitochondrial DNA genealogy. J. Mol. Evol. 35:32-43.

HUDSON, R.R. (1990). Gene genealogies and the coalescent process. Oxford Surveys in Evolutionary Biology 7: 43.

HUDSON, R.R. (1992). Gene trees, species trees and the segregation of ancestral alleles. Genetics 131: 509-512.

HUELSENBECK, J.P. AND D.M. HILLIS. (1993). Success of phylogenetic methods in the four-taxon case. Syst. Biol. 42: 247-264.

HUELSENBECK, J.P., D.L. SWOFFORD, C.W. CUNNINGHAM, J.J. BULL, AND P.J. WADDELL. (1994). Is character weighting a panacea for the problem of data heterogeneity in phylogenetic analysis? Syst. Biol. 43: 288-295.

IOSIFESCU, M. (1980). Finite Markov processes and their applications. John Wiley and sons, New York.

IWABE, N., K.-I. KUMA, M. HASEGAWA, S. OSAWA, AND T. MIYATA. (1989). Evolutionary relationship of archaebacteria, eubacteria, and eukaryotes inferred from phylogenetic trees of duplicated genes. Proc. Natl. Acad. Sci. USA 86: 9355-9359.

JEFFARES, D.C., A.M. POOLE AND D. PENNY. (1995). Pre-rRNA processing and the path from the RNA world. Trends Biochem. Sci. 20: 295-332.

JIN, L. AND M. NEI. (1990). Limitations of the evolutionary parsimony method of phylogenetic analysis. Mol. Biol. Evol. 7: 82-102.

JUKES, T. H., AND C. R. CANTOR. (1969). Evolution of protein molecules. *In* Mammalian Protein Metabolism (ed. H. N. Munro), pp. 21-132. Academic Press, New York.

KABNICK, K.S., AND D.A. PEATTIE. (1991). *Giardia*: A missing link between prokaryotes and eukaryotes. Amer. Sci. 79: 34-43.

KANDLER, O., AND W. ZILLIG, eds.(1986). Archaebacteria '85. Gustav Fischer Verlag, New York.

KARLIN, S., AND H.M. TAYLOR. (1975). A first course in stochastic processes, 2nd edition. Academic press, New York.

KAZAZIAN, H.H., JR., A. CHAKRAVARTI, S.H. ORKIN, AND S.E. ANTONARAKIS. (1983). DNA polymorphisms in the human β-globin gene cluster. *In* Evolution of genes and proteins (ed. M. Nei and R.K. Koehn). Sinauer Assoc., Sunderland, Mass.

KENDALL, M., A. STUART, AND J.K. ORD. (1983). The advanced theory of statistics, Volume 3, "Design and analysis, and time series." Charles Griffin and Company Ltd, London.

KEILSON, J. (1979). Markov chain models - rarity and exponentiality. Applied Mathematical Sciences, Vol. 28. Springer-Verlag, New York.

KIDD, K.K. AND L.L. CAVALLI-SFORZA. (1971). Number of characters examined and error in reconstruction of evolutionary trees, pp. 335-346. *In* Mathematics in the Archaeological and Historical Sciences (ed. F. R. Hodson and P. Tautu). Edinburgh University Press, Edinburgh.

KIDD, K.K. AND L.A. SGARAMELLA-ZONTA. (1971). Phylogenetic analysis: Concepts and methods. Am. J. Hum. Genet. 23: 235-252.

KIM, J. 1993. Improving the accuracy of phylogenetic estimation by combining different methods. Syst. Biol. 42: 331-340.

KIMURA, M. (1980). A simple method for estimating evolutionary rate of base substitutions through comparative studies of nucleotide sequences. J. Mol. Evol. 16:111-120.

KIMURA, M. (1981). Estimation of evolutionary distances between homologous nucleotide sequences. Proc. Natl. Acad. Sci. USA 78: 454-458.

KIMURA, M., AND T. OHTA. (1972). On the stochastic model for estimation of mutational distance between homologous proteins. J. Mol. Evol. 2: 87-90.

KISHINO, H. AND M. HASEGAWA. (1989). Evaluation of the maximum likelihood estimate of the evolutionary tree topologies from DNA sequence data, and the branching order in Hominoidea. J. Mol. Evol. 29: 170-179.

KISHINO, H., AND M. HASEGAWA. (1990). Converting distance to time: An application to human evolution. Methods in Enzymology 183: 550-570.

KISHINO, H., MIYATA, T., AND M. HASEGAWA. (1990). Maximum Likelihood inference of protein phylogeny and the origin of chloroplasts. J. Mol. Evol. 31: 151-160.

KLENK, H.-P., AND W. ZILLIG. (1994). DNA-dependent RNA polymerase subunit B as a tool for phylogenetic reconstruction's: branching topology of the archael domain. J. Mol. Evol. 38: 420-432.

KLENK, H.-P., P. PALM, AND W. ZILLIG. (1994). DNA-dependent RNA polymerases as phylogenetic markers. Syst. Appl. Microbiol. 16: 638-647.

KLUGE, A.G. (1989). A concern for evidence and a phylogenetic hypothesis of relationships among *Epicrates*. (Boidae, Serpentes). Syst. Zool. 38: 7-25.

KOCHER, T.D., AND A.C. WILSON. (1991). Sequence evolution of mitochondrial DNA in humans and chimpanzees: Control region and a protein-coding region. *In* Evolution of Life. (ed. S. Osawa and T. Honjo). Springer-Verlag, Tokyo.

KNOLL, A.H. (1994). Neoproterozoic evolution and environmental change. *In* Early life on Earth, Nobel symposium No. 84 (ed. S. Bengtson), p. 439-449. Columbia University Press, New York.

KRZANOWSKI, W. J. 1988. Principles of multivariate analysis. Oxford: Clarendon Press.

KUHN, T.S. (1977). The essential tension. University of Chicago Press.

KUHNER, M.K., AND J. FELSENSTEIN. (1994). A simulation study of phylogeny algorithms under equal and unequal evolutionary rates. Mol. Biol. Evol. 11: 459-468.

KUMAR, S., K. TAMURA, AND M. NEI. (1993). MEGA: Molecular evolutionary genetics analysis program and manual, v. 1.1. Pennsylvania State University, Pennsylvania.

LAKATOS, I. (1974). Falsification and the methodology of scientific research programs. *In* Criticism and the Growth of Knowledge (ed. I. Lakatos and A.E. Musgrave). Cambridge University Press, Cambridge.

LAKE, J.A. (1986). Eocytes: A new ribosome structure indicates a kingdom with a close relationship to eukaryotes. Proc. Natl. Acad. Sci. USA 81: 3786-3790.

LAKE, J.A. (1987). Rate-independent technique for analysis of nucleic acid sequences: Evolutionary parsimony. Mol. Biol. Evol. 4:167-191.

LAKE, J.A. (1988). Origin of the eukaryotic nucleus determined by rate-invariant analysis of rRNA sequences. Nature 331: 184-186.

LAKE, J.A. (1989). Archaebacterial or eocyte tree. Nature 343: 418-419.

LAKE, J.A. (1991). The order of alignment can bias the selection of tree topology. Mol. Biol. Evol. 8: 378-385.

LAKE, J.A. (1994). Reconstructing evolutionary trees from DNA and protein sequences: Paralinear distances. Proc. Natl. Acad. Sci. USA. 91: 1455-1459.

LANAVE, C., G. PREPARATA, C. SACCONE, AND G. SERIO. (1984). A new method for calculating evolutionary substitution rates. J. Mol. Evol. 20: 86-93.

LÄNGE, S., C. ROZARIO, AND M. MÜLLER. (1994). Primary structure of the hydrogenosomal adenylate kinase of *Trichomonas vaginalis* and its phylogenetic relationships. Mol. Biochem. Parisitol. 66: 297-308.

LECOINTRE, G., H. PHILIPPE, H.L. VÂN LÊ, AND H. LE GUYADER. (1993). Species sampling has a major impact on phylogenetic inference. Mol. Phyl. Evol. 2: 205-224.

LEIPE, D.D., J.H. GUNDERSON, T.A. NERAD, AND M.L. SOGIN. (1993). Small subunit RNA^{+} of Hexamita inflata and the quest for the first branch in the eukaryotic tree. Mol. Biochem. Parisitol. 59: 41-48.

LENTO, G.M., R.E. HICKSON, G.K. CHAMBERS AND D. PENNY. (1995). Use of spectral analysis to test hypotheses on the origin of pinninpeds. Mol. Biol. Evol. 12: 28–52.

LI, W.-H. (1981). A simple method for constructing phylogenetic trees from distance matrices. Proc. Natl. Acad. Sci. USA 78: 1085-1089.

LI, W.-H. AND M. GOUY. (1991). Statistical methods for testing phylogenies. *In* Phylogenetic Analysis of DNA Sequences, p. 249-277 (ed. M.M. Miyamoto and J. Cracraft). Oxford University Press, New York.

LI, W.-H., K.H. WOLFE, J. SOURDIS, AND P.M. SHARP. (1987). Reconstruction of phylogenetic trees and estimation of divergence times under non-constant rates of evolution. Cold Spring Harbor Symposia on Quantitative Biology, LII: 847-856.

LI, W.-H. AND A. ZHARKIKH. (1995). Statistical tests of DNA phylogenies. Syst. Biol. 44: 49-63.

LOCKHART, P.J. (1990). Inference of green chloroplast origins. Ph.D. thesis, Sydney University.

LOCKHART, P.J., C.J. HOWE, D.A. BRYANT, T.J. BEANLAND, AND A.W.D. LARKUM. (1992a). Substitutional bias confounds inference of cyanelle origins from sequence data. J. Mol. Evol. 34: 153-162.

LOCKHART, P. J., A. W. LARKUM, P. J. WADDELL, M. A. STEEL, AND D. PENNY. (1995). Evolution of chlorophyll and bacteriochlorophyll: The problem of invariant sites in sequence analysis. Proc. Natl. Acad. Sci. USA (in press).

LOCKHART, P.J., D. PENNY, M.D. HENDY, C.J. HOWE, T.J. BEANLAND, AND A.W.D. LARKUM. (1992b). Controversy on chloroplast origins. FEBS Lett. 301: 127-131.

LOCKHART, P.J., M.A. STEEL, M.D. HENDY AND D. PENNY. (1994). Recovering evolutionary trees under a more realistic model of sequence evolution. Mol. Biol. Evol. 11: 605-612.

LOOMIS, W.F., AND D.W. SMITH. (1990). Molecular phylogeny of *Dictyostelium discoideum* by protein sequence comparison. Proc. Natl. Acad. Sci. USA 87: 9093-9097.

MCCULLAGH, P., AND J.A. NELDER. (1983). Generalized linear models. Chapman and Hall, London.

MCCULLAGH, P., AND J.A. NELDER. (1989). Generalised linear models, second edition. Chapman and Hall, London.

MCDADE, L. (1990). Hybrids and phylogenetic systematics I. Patterns of character expression in hybrids and their implications for cladistic analysis. Evolution 44: 1685-1700.

MCDADE, L. (1992). Hybrids and phylogenetic systematics II. The impact of hybrids on cladistic analysis. Evolution 46: 1329-1346.

MADDISON, W.P., AND D.R. MADDISON. (1992). MacClade, version 3. Sinauer Associates, Sunderland, Massachusetts.

MADDISON, D.R., M. RUVOLO AND D.L. SWOFFORD. (1992). Geographic origins of human mitochondrial DNA: Phylogenetic evidence from control region sequences. Syst. Biol. 41: 111-124.

MAHER, P. (1993). Betting on theories. Cambridge University Press, Cambridge.

MANLY, B.F.J. (1986). Multivariate statistical methods: A primer. Chapman and Hall, London.

MARSHALL, C.R. (1990). The fossil record and estimating divergence times between lineages: Maximum divergence times and the importance of reliable phylogenies. J. Mol. Evol. 30: 400-408.

MILLER, A.J. (1990). Subset selection in regression. Chapman and Hall, London.

MINOUX, M. (1986). Mathematical programming; theory and algorithms. John Wiley and sons, New York.

MIYAMOTO, M.M., J.L. SLIGHTOM AND M. GOODMAN. (1987). Phylogenetic relationships of humans and African apes as ascertained from DNA sequences (7.1 kilobase pairs) of the φη-globin region. Science 238: 369-373.

MORIN, P.A., J.J. MOORE, R. CHAKRABORTY, L. JIN, J. GOODALL, AND D.S. WOODRUFF (1994). Kin selection, social structure, gene flow, and the evolution of chimpanzees. Science 265: 1193-1201.

MUSE, S. V., AND B. S. GAUT. (1994). A likelihood approach for comparing synonymous and nonsynonymous substitution rates, with application to the chloroplast genome. Mol. Biol. Evol. 11:715-724.

NAVIDI, W.C., AND L. BECKETT-LEMUS. (1992). The effect of unequal transversion rates on the accuracy of evolutionary parsimony. Mol. Biol. Evol. 9: 1163-1175.

NAVIDI, W.C., G.A. CHURCHILL AND A. VON HAESELER. (1991). Methods for inferring phylogenies from nucleic acid sequence data by using maximum likelihood and linear invariants. Mol. Biol. Evol. 8: 128-143.

NEI, M. (1987). Molecular evolutionary genetics. Columbia University Press, New York.

NEI, M., and L. JIN. (1989). Variances of the average numbers of nucleotide substitutions within and between populations. Mol. Biol. Evol. 6: 290-300.

OLSEN, G.J. (1987). The earliest phylogenetic branchings: Comparing rRNA-based evolutionary trees inferred with various techniques. Cold Spring Harbor Symp. Quant. Biol. 52: 825-837.

OLSEN, G.J. (1994). DNArates (a computer program in C source code). Available by anonymous ftp from rdp.life.uiuc.edu (in directory pub/RDP/programs/fastDNAml).

OLSEN, G.J., AND C.R. WOESE. (1989). A brief note concerning archaebacterial phylogeny. Canadian Journal of Microbiology 35: 119-123.

OLSEN, G.J., C.R. WOESE. (1993). Ribosomal RNA: a key to phylogeny. FASEB J 7: 113-123.

OLSEN, G.J., C.R. WOESE, AND R. OVERBEEK. (1993). The winds of (evolutionary) change: Breathing new life into microbiology. Journal of Bacteriology 176: 1-6.

OLSEN, G.J., R. OVERBEEK, N. LARSEN, T.L. MARSH, M.J. MCCAUGHEY, M.A. MACIUKENAS, W.-M. KUAN, T.J. MACKE, Y. XING AND C.R. WOESE. (1992). The ribosomal database project. Nucleic Acids Research 20 (suppl.):2100-2200.

PEARL, J., AND M. TARSI. (1986). Structuring casual trees. Journal of Complexity, 2: 60-77.

PENNY, D. (1982). Towards a basis for classification: The incompleteness of distance measures, incompatibility analysis and phenetic classification. J. Theor. Biol. 96:129–142.

PENNY, D. AND M.D. HENDY. (1985). Testing methods of evolutionary tree construction. Cladistics 1: 266-272.

PENNY, D., AND M.D. HENDY. (1986). Estimating the reliability of evolutionary trees. Mol. Biol. Evol. 3: 403-417.

PENNY, D., AND M.D. HENDY. (1987). TurboTree: A fast algorithm for minimal trees. Computer Applications in the Biosciences 3: 183-188.

PENNY, D., M.D. HENDY, AND M.A. STEEL. (1991). Testing the theory of descent. *In* Phylogenetic analysis of DNA sequences (ed. M.M. Miyamoto and J. Cracraft), p. 155-183. Oxford University press, New York.

PENNY, D., M.D. HENDY, AND I.M. HENDERSON. (1987). The reliability of evolutionary trees. Cold Spring Harbor Symp. Quant. Biol. 52: 857-862.

PENNY, D., M. D. HENDY AND M. A. STEEL. (1992). Progress with methods for constructing evolutionary trees. Trend. Ecol. Evol. 7: 73-79.

PENNY, D., M.D. HENDY, E.A. ZIMMER, AND R.K. HAMBY (1990). Trees from sequences: Panacea or Pandora's box? Australian Journal of Botany 3: 21-38.

PENNY, D., P.J. LOCKHART, M.A. STEEL, AND M.D. HENDY. (1994). The role of models in reconstructing evolutionary trees. *In* Models in phylogeny reconstruction (ed. R.W. Scotland, D.J. Siebert, and D.M. Williams), p. 211-229. Clarendon Press, Oxford.

PENNY, D., M.A. STEEL, P.J. WADDELL AND M.D. HENDY. (1995). Improved analyses of human mtDNA sequence support a recent African origin for *Homo sapiens*. Mol. Biol. Evol. 12: 863-882.

PENNY, D., E.E. WATSON, R.E. HICKSON, AND P.J. LOCKHART. (1993). Some recent progress with methods for evolutionary trees. New Zealand Journal of Botany (Conference Issue) 31: 275-288.

PERRIN-PECONTAL, P., M. GOUY, V.-M. NIGON, AND G. TRABUCHET. (1992). Evolution of the primate β-globin gene region: Nucleotide sequence of the δ–β-globin intergenic region of gorilla and phylogenetic relationships between African apes and man. J. Mol. Evol. 34: 17-30.

PHILLIPS, P.C.B. (1982). The true characteristic function of the F distribution. Biometrika 69:261-264.

POPPER, K.R. (1979). Objective knowledge: An evolutionary approach (revised edition). Clarendon Press, Oxford.

READ, T.R.C., AND N.A.C. CRESSIE. (1988). Goodness-of-fit statistics for discrete multivariate data. Springer-Verlag, New York.

REEVES, J.H. (1992). Heterogeneity in the substitution process of amino acid sites of proteins coded for by Mitochondrial DNA. J. Mol. Evol. 35: 17-31.

RÉNYI, A. (1970). Probability theory. North Holland publishing, Amsterdam.

RICE, W.R. (1989). Analyzing tables of statistical tests. Evolution 43: 223-225.

RITLAND, K., AND M.T. CLEGG. (1987). Evolutionary analysis of plant DNA sequences. The American Naturalist 130, supplement: S74-S100.

RIVERA, M.C., AND J.A. LAKE. (1992). Evidence that eukaryotes and eocyte prokaryotes are immediate relatives. Science. 257: 74-76.

ROBINSON, D.F., AND L.R. FOLDS. (1979). Comparison of weighted labeled trees. Lecture notes in mathematics, 748: 119-126. Springer-Verlag, Berlin.

ROBINSON, D.F., AND L.R. FOLDS. (1981). Comparison of phylogenetic trees. Mathematical Biosciences 53: 131-147.

ROGERS, A.R., AND H. HARPENDING. (1992). Population growth makes waves in the distribution of pairwise genetic differences. Mol. Biol. Evol. 9: 552-569.

RODRIGO, A.G. (1993). Calibrating the bootstrap test of monophyly. International Journal for Parasitology 23: 507-514.

RODRIGO, A.G., AND P.R. BERGQUIST, and P.L. BERGQUIST. (1994). Inadequate support for an evolutionary link between the Metazoa and the Fungi. Syst. Biol. 43: 578-584.

RODRIGO, A.G., M. KELLY-BORGES, P.R. BERGQUIST AND P.L. BERGQUIST. (1993). A randomisation test of the null hypothesis that two cladograms are sample estimates of a parametric phylogenetic tree. New Zealand Journal of Botany (Conference Issue) 31: 257-268.

RODRÍGUEZ, F., AND J.R. MEDINA. (Feb. 1986, unpublished manuscript). Solution of the general stochastic model of nucleotide substitution (rejected by Journal of Molecular Evolution).

RODRÍGUEZ, F., J.L. OLIVER, A. MARIN AND J.R. MEDINA. (1990). The general stochastic model of nucleotide substitution. J. Theor. Biol. 142: 485-501.

ROFF, D. A. AND P. BENTZEN. (1989). The statistical analysis of mitochondrial DNA polymorphisms: χ^2 and the problem of small samples. Mol. Biol. Evol. 6: 539-545.

ROGERS, A.R., AND H. HARPENDING. (1992). Population growth makes waves in the distribution of pairwise genetic differences. Mol. Biol. Evol. 9: 552-569.

ROM, D.R. (1990). A sequentially rejective test procedure based on a modified Bonferroni inequality. Biometrika 77: 663-665.

RZHETSKY, A., and M. NEI. 1992a. A simple method for estimating and testing minimum-evolution trees. Mol. Biol. Evol. 9: 945-967.

RZHETSKY, A., and M. NEI. 1992b. Statistical properties of the ordinary least-squares, generalised least-squares, and minimum-evolution methods of phylogenetic inference. J. Mol. Evol. 35: 367-375.

RZHETSKY, A., and M. NEI. 1993. Theoretical foundation of the minimum-evolution method of phylogenetic inference. Mol. Biol. Evol. 10: 1073-1095.

RZHETSKY, A., AND M. NEI. (1995). Tests of applicability of several models for DNA sequence data. Mol. Biol. Evol. 12: 131-151.

SACCONE, C., G. PESOLE, AND G. PREPARATA. (1989). DNA microenvironments and the molecular clock. J. Mol. Evol. 29: 407-411.

SACCONE, C., C. LANAVE, G. PESOLE, AND G. PREPARATA. (1990). Influence of base composition on quantitative estimates of gene evolution. Methods in Enzymology 183: 584-598.

SAITOU, N. (1990). Maximum likelihood methods. Meth. Enzymol. 183: 584-598.

SAITOU, N. AND M. NEI. (1987). The neighbor-joining method: A new method for reconstructing phylogenetic trees. Mol. Biol. Evol. 183: 584-598.

SAKAMOTO, Y., M. ISHIGURO, AND G. KITAGAWA. (1986). Akaike information criterion statistics. KTK Scientific Publishers, Tokyo.

SANDERSON, M.J. (1989). Confidence limits on phylogenies: The bootstrap revisited. Cladistics 5: 113-129.

SANKOFF, D. (1975). Minimal mutation trees of sequences. SIAM J. Appl. Math. 28:35-42.

SANKOFF, D., R.J. CEDERGREN, AND G. LAPALME. (1976). Frequency of insertion-deletion, transversion, and transition in the evolution of 5S ribosomal RNA. J. Mol. Evol. 7:133-149.

SATTATH, S., AND A. TVERSKY. (1977). Additive similarity trees. Psychometrika 42: 319-345.

SCHÖNIGER, M., AND A. VON HAESELER. (1993). A simple method to improve the reliability of tree reconstructions. Mol. Biol. Evol. 10: 471-483.

SCHÖNIGER, M., AND A. VON HAESELER. (1994). A stochastic model for the evolution of autocorrelated DNA sequences. Mol. Phyl. Evol. 3: 240-247.

SEBER, G.A.F. (1982). The estimation of animal abundance. Charles Griffin and Co., London.

SHOEMAKER, J.S., AND W.M. FITCH. (1989). Evidence from nuclear sequences that invariable sites should be considered when sequence divergence is calculated. Mol. Biol. Evol. 6: 270-289.

SIBLEY, C.G., J.A. COMSTOCK AND J.E. AHLQUIST. (1990). DNA hybridization evidence of Hominoid phylogeny: A reanalysis of the data. J. Mol. Evol. 30: 202-236.

SIDDALL, M.E., H. HONG, AND S.S. DESSER. (1992). Phylogenetic analysis of the Diplomonadida (Wenyon, 1926) Brugerolle, 1975: Evidence for heterochrony in protozoa and against *Giardia lamblia* as a "missing link". The Journal of Protozoology 39: 361-367.

SIDOW, A., T. NGUYEN, AND T.P. SPEED. (1992). Estimating the fraction of invariable codons with a capture-recapture method. J. Mol. Evol. 35: 253-260.

SIDOW, A., AND W.K. THOMAS. (1994). A molecular evolutionary framework for eukaryotic model organisms. Current Biology 4: 596-603.

SMITH, C.A.B. (1993). Review of "Likelihood, expended edition" (By A.W.F. Edwards). Ann. Hum. Genet. 57: 315-321.

SOGIN, M.L. (1989). Evolution of microorganisms and their small subunit ribosomal RNA's. Am. Zool. 29: 487-499.

SOGIN, M. (1991). Early evolution and the origins of eukaryotes. Current Opinion in Genetics and Development 1: 457-463.

SOGIN, M.L., J.H. GUNDERSON, H.J. ELWOOD, R.A. ALONSO, D.A. PEATTIE. (1989). Phylogenetic meaning of the kingdom concept: An unusual ribosomal RNA from *Giardia lamblia*. Science 243: 75-77.

SOGIN, M.L., G. HINKLE, AND D.D. LEIPE. (1993). Universal tree of life. Nature 362: 795.

SOKAL, R.R., AND F.J. ROHLF. (1981). Biometry, second edition. W.H. Freeman and Co., San Francisco.

SPRAGUE, V. (1977). Classification and phylogeny of the Microsporidia. *In* "Comparative Pathobiology, Vol. 2, Systematics of the Microsporidia" (ed. L.A. Bulla and T.C. Cheng). Plenum, New York.

STEEL, M.A. (1989). Distributions on bicoloured evolutionary trees. PhD. Thesis, Massey University.

STEEL, M.A. (1994a). Recovering a tree from the leaf colourations it generates under a Markov model. Appl. Math. Lett. 7: 19-23.

STEEL, M.A. (1994b). The maximum likelihood point for a phylogenetic tree is not unique. Syst. Biol. 43:560-564.

STEEL, M.A., M.D. HENDY, AND D. PENNY (1988). Loss of information in genetic distances. Nature 336: 118.

STEEL, M.A., M. D. HENDY AND D. PENNY. (1993a). Invertible models of sequence evolution. Mathematical and Information Sciences Report, Preprint Series: 93/02, Massey University, Palmerston North, New Zealand.

STEEL, M.A., M.D. HENDY AND D. PENNY. (1993b). Parsimony can be consistent! Syst. Biol. 42: 581-587

STEEL, M.A., M.D. HENDY, L.A. SZÉKELY and P.L. ERDÖS. (1992). Spectral analysis and a closest tree method for genetic sequences. Appl. Math. Lett. 5: 63-67.

STEEL, M.A., P.J. LOCKHART AND D. PENNY (1993d). Confidence in evolutionary trees from biological sequence data. Nature 364: 440-442.

STEEL, M.A., AND D. PENNY. (1993). Distributions of tree comparison metrics -some new results. Syst. Biol. 42: 126-141.

STEEL, M.A., L. SZÉKELY, P.L. ERDÖS AND P.J. WADDELL. (1993c). A complete family of phylogenetic invariants for any number of taxa under Kimura's 3ST model. New Zealand Journal of Botany (Conference Issue) 31: 289-296.

STEEL, M.A., L.A. SZÉKELY, AND M.D. HENDY. (1994e). Reconstructing trees when sequence sites evolve at variable rates. J. Comp. Biol. 1: 153-163.

STEPHENS, J.C. (1985). Statistical methods of DNA sequence analysis: Detection of intragenic recombination or gene conversion. Mol. Biol. Evol. 2: 539-556.

STEWART, C.-B. AND A.C. WILSON. (1987). Sequence convergence and functional adaptation of stomach lysozymes from foregut fermenters. Cold Spring Harbor Symp. Quant. Biol. 52: 891-899.

STUART, A., and J.K. ORD. (1987). Kendall's advanced theory of statistics. Volume 1. 5th ed. Edward Arnold, London.

STUART, A., and J.K. ORD. (1990). Kendall's advanced theory of statistics. Volume 2, Distribution theory. Classical inference and relationship. 5th ed. Edward Arnold, London.

STUDIER, J.A., AND K.J. KEPPLER. (1988). A note on the neighbor-joining algorithm of Saitou and Nei. Mol. Biol. Evol. 5: 729-731.

SWOFFORD, D.L. (1993). PAUP: Phylogenetic Analysis Under Parsimony, version 3.1. Computer program and printed manual. Formerly distributed by the Illinois Natural History Survey, Champaign, Illinois.

SWOFFORD, D.L. (1995). *PAUP**, version 4.0. Currently test versions of computer program, but to be published by: Sinauer Associates, Sunderland, Massachusetts.

SWOFFORD, D.L., AND G.J. OLSEN. (1990). Phylogeny reconstruction. In Molecular Systematics, (ed. D. M. Hillis and C. Moritz), pp 411-501. Sinauer Associates, Sunderland, Massachusetts.

SWOFFORD, D.L., G.J. OLSEN, P.J. WADDELL, AND D.M. HILLIS. (1995). Phylogenetic Inference. *In* Molecular Systematics, second edition (ed. D. M. Hillis, C. Moritz and B.K. Marble). Sinauer Associates, Sunderland, Massachusetts (in press).

SZÉKELY, L.A., M.A. STEEL, AND P.L. ERDÖS. (1993). Fourier calculus on evolutionary trees. Advances in Applied Mathematics 14: 200-216.

TAJIMA, F. (1993a). Unbiased estimation of evolutionary distance between nucleotide sequences. Mol. Biol. Evol. 10: 677-688.

TAJIMA, F. (1993b). Simple methods for testing the molecular evolutionary clock hypothesis. Genetics 135: 599-607.

TAJIMA, F., AND M. NEI. (1982). Biases of the estimates of DNA divergence obtained by the restriction enzyme technique. J. Mol. Evol. 18: 115-120.

TAKAHATA, N. (1993). Allelic genealogy and human evolution. Mol. Biol. Evol. 10:2-22.

TAMURA, K. (1992). Estimation of the number of nucleotide substitutions when there are strong transition-transversion and G+C-content biases. Mol. Biol. Evol. 9: 678-687.

TAVARÉ, S. (1986). Some probabilistic and statistical problems in the analysis of DNA sequences. Lectures on Mathematics in the Life Sciences 17: 57-86.

TEMPLETON, A.R. (1983). Phylogenetic inference from restriction endonuclease cleavage site maps with particular reference to the evolution of humans and apes. Evolution 37: 221-244.

TILLIER, E.R.M. (1994). Maximum likelihood with multiparameter models. J. Mol. Evol. 39: 409-417.

TILLIER, E.R.M., AND R.A. COLLINS. (1995). Neighbor joining and maximum likelihood with RNA sequences: Addressing the interdependence of sites. Mol. Biol. Evol. 12: 7-15.

TOLIMIERI, R., M. AN and C. LU. 1989. Algorithms for discrete Fourier transform and convolution. Springer-Verlag, New York.

UZZEL, T., AND K. W. CORBIN. (1971). Fitting discrete probability distributions to evolutionary events. Science 172: 1089-1096.

VAN DE PEER, Y., J.-M. NEEFS, P. DE RIJK, AND R. DE WACHTER. (1993). Reconstructing evolution from eukaryotic small-ribosomal-subunit RNA sequences: Calibration of the molecular clock. J. Mol. Evol. 37: 221-232.

VAN DE PEER, Y., J.-M. NEEFS, P. DE RIJK, P. DE VOS, AND R. DE WACHTER. (1994). About the order of divergence of the major bacterial taxa during evolution. System. Appl. Microbiol. 17: 32-38.

VAWTER, L., AND W.M. BROWN. (1993). Rates and patterns of base change in the small subunit ribosomal RNA gene. Genetics 134: 597-608.

VIGILANT, L., M. STONEKING, H. HARPENDING, K. HAWKES AND A.C. WILSON (1991). African populations and the evolution of human mitochondrial DNA. Science 253: 1503-1507.

VOSSBRINCK, C.R., M.D. BAKER, E.S. DIDIER, B.A. DEBRUNNER-VOSSBRINK, AND J.A. SHADDUCK. (1986). Ribosomal DNA sequences of *Encephalitozoon hellem* and *Encephalitozoon cuniculi*: Species identification and phylogenetic construction. J. Euk. Mircobiol. 40: 354-362.

VOSSBRINCK, C.R., AND C.R. WOESE. (1986). Eukaryotic ribosomes that lack a 5.8S RNA. Nature 320: 287-288.

VON HAESELER, A., AND G. A. CHURCHILL. (1993). Network models for sequence evolution. J. Mol. Evol. 37: 77-85.

WADDELL, P.J. (1990). The mating behaviour and phylogeny of the *Nasuta* subgroup of *Drosophila*. MSc. Thesis, University of Auckland.

WADDELL, P.J. AND D. PENNY. (1995). Evolutionary trees of apes and humans from DNA sequences. *In* Handbook of Human Symbolic Evolution. (ed. A.J. Lock and C.R. Peters). Clarendon Press, Oxford. (in press since August 1993, preprints available).

WADDELL, P.J., D. PENNY, M.D. HENDY AND G. ARNOLD. (1994). The sampling distributions and covariance matrix of phylogenetic spectra. Mol. Biol. Evol.: 630-642.

WADDELL, P.J., AND M.A. STEEL. (1995). Time reversible distances allowing a distribution of rates across sites. Department of Mathematics and Statistics Report, University of Canterbury (in preparation).

WAINRIGHT, P.O., HINKLE, M.L. SOGIN, AND S.K. STICKEL. (1993). The monophyletic origins of the Metazoa; an unexpected evolutionary link with Fungi. Science 260, 340-243.

WAKELEY, J. (1993). Substitution rate variation among sites in hypervariable region 1 of human mitochondrial DNA. J. Mol. Evol. 37: 613-623.

WEISBURG, W.G., GIOVANNONI, S.J., AND C.R. WOESE. (1989). The *Deinococcus-Therums* phylum and the effect of rRNA composition on phylogenetic tree reconstruction. System. Appl. Microbiol. 11: 128-134.

WHITE, T.D., G. SUWA, AND B. ASFAW. (1994). *Australopithecus ramidus*, a new species of early hominid from Aramis, Ethiopia. Nature 371: 306-312.

WILLIAMS, P. L. AND W. M. FITCH. (1989). Finding the minimal change in a given tree, pp. 453–470. *In* The Hierarchy of Life (ed. B. Fernholm, K. Bremer and H. Jörnvall). Elsevier, Amsterdam.

WOESE, C.R., L. ACHENBACH, P. ROUVIERE, AND L. MANDELCO. (1991). Archaeal phylogeny: Reexamination of the phylogenetic position of *Archaeoglobus fulgidus* in light of certain composition-induced artifacts. System. Appl. Microbiol. 14: 364-371.

WOESE, C.R., AND G.E. FOX. (1977). Phylogenetic structure of the prokaryotic domain: The primary kingdoms. Proc. Natl. Acad. Sci. USA, 74: 5088-5090.

WOESE, C.R. AND G.J. OLSEN. (1986). Archaebacterial phylogeny: Perspectives on the urkingdoms. Syst. Appl. Microbiol. 7: 161-177.

WOESE, C.R., AND R.S. WOLFE, eds. (1985). The Bacteria: Volume 8, Archaebacteria. Academic Press, Orlando, Florida.

WOLTERS, J., AND V.A. ERDMANN. (1986). Cladistic analysis of 5S rRNA and 16S rRNA secondary and primary structure - The evolution of eukaryotes and their relation to archaebacteria. J. Mol. Evol. 24: 152

WU, C.-I. (1991). Inferences of species phylogeny in relation to segregation of ancient polymorphisms. Genetics 127: 429-435.

YANG, Z. (1993). Maximum-likelihood estimation of phylogeny from DNA sequences when substitution rates differ over sites. Mol. Biol. Evol. 10: 1396-1401.

YANG, Z. (1994). Maximum Likelihood phylogenetic estimation from DNA sequences with variable rates over sites: Approximate methods. J. Mol. Evol. 39: 306-314.

YANG, Z., N. GOLDMAN, AND A. FRIDAY. (1994). Comparison of models of nucleotide substitution used in maximum-likelihood estimation. Mol. Biol. Evol. 11: 316-324.

YANG, Z., AND D. ROBERTS. (1995). On the use of nucleic acid sequences to infer early branchings in the tree of life. Mol. Biol. Evol. 12: 451-458.

ZHARKIKH, A. (1994). Estimation of evolutionary distances between nucleotide sequences. J. Mol. Evol. 39:315-329.

ZHARKIKH, A., AND W.-H. LI. (1992a) Statistical properties of bootstrap estimation of phylogenetic variability from nucleotide sequences. I. Four taxa with a molecular clock. Mol. Biol. Evol. 9: 1119-1147.

ZHARKIKH, A. AND W.-H. LI. (1992b). Statistical properties of bootstrap estimation of phylogenetic variability from nucleotide sequences. II. Four taxa without a molecular clock. J. Mol. Evol. 35:356-366.

ZHARKIKH, A. AND W.-H. LI. (1995). Estimation of confidence in phylogeny: The full-and-partial bootstrap technique. Mol. Phylogenet. Evol. 4:44-63.

ZUCKERKANDL, E., AND L. PAULING. (1965). Molecules as documents of evolutionary history. Journal of theoretical biology. 8: 357-366.