THE COVARION MODEL OF MOLECULAR EVOLUTION

BENNET M^{c}COMISH

1997

THE COVARION MODEL OF MOLECULAR EVOLUTION

A thesis presented in partial fulfilment

of the requirements for the degree of

Master of Philosophy in Biology

at Massey University

Bennet M^{c}Comish

1997

Abstract

Current methods for constructing evolutionary trees generally do not work well for sequences in which multiple substitutions have occurred. The covarion hypothesis may provide a solution to this problem. This hypothesis states that only a limited number of the codons in a given sequence are free to vary, but that the set of variable codons may change as mutations are fixed in the population. Although this is reasonable from a biological point of view, it is a difficult hypothesis to test scientifically because the apparent large number of parameters involved makes it very hard to analyse statistically.

In this study, computer simulations were carried out on up to 51 machines running in parallel, using a simple covarion model based on a hidden Markov model (HMM) approach. This model required two new parameters—the proportion of sites that are variable at any given time, and the rate of exchange between fixed and variable states. These two parameters were both varied in the simulations. Sequence and distance data were simulated on a given tree under this covarion model, and these data were used to test the performance of standard tree-building methods at recovering the original tree

The neighbour joining and maximum likelihood methods tested were found to perform better with data generated under the covarion model than with data generated under a simpler model in which all sites vary at the same rate. This suggests that current tree-building methods may perform better with biological data than computer simulation studies suggest.

Acknowledgements

I am deeply grateful to everyone who has helped me, directly or indirectly, to get through this thesis.

I am most indebted to my supervisor, David Penny, without whose wise and patient guidance I would not have been able to complete this work. His ideas and programming skills were also crucial to the project.

I would also like to thank Mike Steel, who came up with the initial idea of using a hidden Markov model for the covarion hypothesis.

Various people here in the Department of Plant Biology and Something Else, and elsewhere, also deserve thanks for their ideas and other helpful things. These include Dan Jeffares, Anthony Poole, Kerryn Slack, Peter Waddell and Abby Harrison for numerous interesting discussions (in most cases completely unrelated to this project, but valuable none the less), Mike Hendy, Mike Steel, Pete Lockhart, Mike Charleston, Chris Tuffley and others for less frequent discussions of more immediate relevance, and Ted Drawneek for his assistance in setting up the framework for running simulations remotely and in parallel.

I thank all my friends (some of whom are mentioned above, but also Johan van Beek, Paul Hirst, Matt Barclay, and too many others to name here) for altering my sanity levels in whichever direction was necessary at the time, and not allowing me to stay in touch with reality for too much of the time.

I am eternally grateful to the various organisations to which I have sold my Eternal Soul™, and from which I have purchased Eternal Salvation™.

Finally, I thank both my parents for their support during this sometimes rather trying period of my life.

Table of Contents

List of Figures

List of Tables

List of Abbreviations and Symbols

2ST	Kimura two substitution-type model
3ST	Kimura three substitution-type model
covarion	Concomitantly variable codon
DNA	Deoxyribonucleic acid
HMM	Hidden Markov model
i.i.d.	Independent and identically distributed
SOD	Cu,Zn superoxide dismutase

Parameters used in the simulations:

α	Rate of transitions (A↔G and C↔T) in the 3ST model
β	Rate of type 1 transversions (A↔T and C↔G) in the 3ST model
γ	Rate of type 2 transversions (A↔C and G↔T) in the 3ST model
c	Sequence length
e	Rate of exchange between fixed and variable states
f	Proportion of sites that are permanently fixed
n	Number of taxa
r	Number of character states
s	Number of simulations to be performed
v	Proportion of sites that are variable

1. Introduction

1.1 Overview of the problem

A phylogenetic tree represents a hypothesis concerning the relationship between a set of taxa. Trees are constructed so that inferences can be made about the biology of the taxa. Tree-building methods can either use molecular sequence data directly, or they can use distances or dissimilarities between taxa.

Because events early in the history of life are not directly observable, it is difficult to know the accuracy of these evolutionary trees. Multiple substitutions at a site are an important source of error in the reconstruction of phylogenetic trees from sequence data, particularly when looking at older divergences. For example, if a sequence evolves at a rate of 3.3×10^{-9} substitutions per site per year (a realistic rate, Bulmer *et al.*, 1991), then after a billion years 5 changes would be expected to have occurred at each site in each lineage.

Some sites in the sequence may be conserved for functional reasons and do not change at all over time. Apart from assisting with alignment, these do not give information that can be used in selecting an optimal tree. At the other extreme, multiple substitutions can result in randomisation of the sequence at the sites that are free to vary, so that for ancient divergences these sites no longer contain information that can be used to infer relationships between the sequences. Computer simulations have shown that current methods are very unreliable when there has been an average of more than one change per site (Charleston, 1994, p.139).

Thus we appear to have a paradox. For sequences that diverged over a billion years ago, we expect sites to be either constant (and contain no relevant information), or so variable that they are saturated (and no longer contain any relevant information). The aim of this project is to investigate the covarion hypothesis as a means of finding a way out of this paradox.

In order to correct sequences for multiple substitutions we must assume a mechanism of sequence evolution. This mechanism will usually consist of a transition matrix determining the probability that nucleotide i at site x changes to nucleotide j along a given edge of the tree.

Many mechanisms have been suggested to describe sequence evolution, and these often assume that changes in the sequence are 'independent and identically distributed'(Penny *et al.*, 1992). That is, they assume all sites follow the same underlying process of substitution (identically distributed), and that a change at one site does not affect the chances of a change occurring at any other site, or in any other lineage (independent). These are known as i.i.d. models.

Commonly used models of sequence evolution include the Jukes-Cantor model (Jukes and Cantor, 1969) and the Kimura two and three substitution-type (2ST and 3ST) models (Kimura, 1980 and 1981 respectively). These models assume that changes are independent and identically distributed, and that the mechanism of change is constant over the whole tree. The Jukes-Cantor model assumes that all the rates of substitution from one base to another are equal, while Kimura's 2ST model allows for two rates, α and β, one each for transitions and transversions. The 3ST model is more general again, allowing two parameters, β and γ, for transversions (β for A$\leftrightarrow$T and G$\leftrightarrow$C; γ for A$\leftrightarrow$C and G$\leftrightarrow$T). Because all three models are symmetric (the rates of change in both directions between any two nucleotides, say A$\rightarrow$G and G$\rightarrow$A, are equal), the mean frequency of each of the four bases is expected to be the same and eventually end up as 1/4.

The 'identically distributed' assumption in most cases is not true, since it has been shown on many occasions that different sites in a macromolecule evolve at different rates (Fitch, 1971a; Uzzell and Corbin, 1971; Holmquist *et al.*, 1983; Tajima and Nei, 1984). There are two reasons for different rates of evolution at different sites—they may have either different mutation rates, or different selective constraints. The 'hot spots' observed in the D-loop of mitochondrial DNA (Wakeley, 1993) are an example of the former, but these

may be uncommon, at least in sequences used in phylogenetic analyses. Different selective constraints at different sites are likely to be much more common. In proteins, amino acid residues involved in the active site tend to evolve more slowly than other sites in the molecule, as predicted by Kimura's Neutral Theory of molecular evolution (Kimura, 1983), while for protein-coding DNA the third positions of codons evolve faster than the first and second.

Some models do take into account different rates for different codon positions, but still assume that changes are independent. However, a substitution at one site may alter the selective constraints on some other sites, allowing further substitutions, so in a formal sense changes are not independent. It has been shown that models that ignore correlated sites underestimate the actual amount of divergence (Schöniger and von Haeseler, 1994).

The covarion hypothesis, first proposed by Fitch and Markowitz in 1970, suggests a much more realistic, but also much more complex, mechanism. The complexity of this mechanism, however, makes it very difficult to analyse mathematically. This study uses a simplified covarion-style model as a basis for computer simulations in order to test the performance of tree-building methods on more realistic data than that produced by more widely used models.

1.2 The covarion hypothesis

Fitch and Markowitz (1970) observed that when 29 cytochrome *c* sequences from fungi, plants and animals were compared, 32 of the 113 codons were constant over all 29 taxa. When they reduced the range of species to non-primate mammals, however, the number of codons that were invariant rose to 95. They explained this by postulating that

> "because of the structural restraints imposed by functional requirements, mutations that will not be selected against are available only for a very limited number of positions. We shall use the term acceptable for such mutations. However, as such acceptable mutations are fixed they alter the positions in which other acceptable mutations may be fixed. Thus, only

about ten codons, on the average, in any cytochrome *c* may have acceptable mutations available to them but the particular codons will vary from one species to another. We shall term those codons at any one instant in time and in any given gene for which an acceptable mutation is available as the *concomitantly variable codons.*[*] "

This means that although most codons in a gene may be found to vary over a wide range of species, very few of the codons in a given species may be free to vary at a given point in time.

Fitch and Markowitz suggested that the interdependence of events at different coding positions may be related to the observations of Wyckoff (1968), who noted a spatial relationship between pairs of substitutions observed between rat and bovine ribonucleases.

In general, a mutation at one site in a protein is likely to alter the constraints on the molecule, so that some sites become free to change and others are no longer free to change. The sites affected in this way are likely to be close to the site in which the mutation occurs, that is, close in terms of the three-dimensional structure of the protein, not necessarily close in the sequence. This idea can be extended to the external constraints on the protein, so that a mutation in one protein may alter the covarion set of other proteins with which it interacts.

The covarion concept has also been extended to that of *co*ncomitantly *vari*able nucle*otides* or covariotides (Fitch, 1986), and applied to RNA, where the interactions between sites are mainly limited to complementary base pairing.

---

[*] The term 'covarions' was coined as an abbreviation of the phrase concomitantly variable codons.

1.3 History of the covarion hypothesis

In the two and a half decades since the covarion hypothesis was first proposed, only a handful of studies have analysed it in any detail. Notable examples of these detailed studies are Fitch (1971a), Karon (1979), Fitch and Ayala (1994), and Miyamoto and Fitch (1995).

Fitch (1971a) used a mathematical model for the covarion hypothesis to estimate the number of covarions, c, in cytochrome *c*, and the persistence of variability, v, of the covarions (where the persistence of variability is the probability of any given covarion retaining its variable status after a substitution elsewhere in the gene). This was done by examining the rate at which observable double mutations occur on a phylogenetic tree, and comparing this to expected values calculated for given values of c and v. The best fit was found to be between 4 and 10 covarions, with a persistence of variability of less than 0.25. This means that for each mutation fixed, 75% or more (on average) of the covarions lose their variable status.

Karon (1979) improved Fitch's mathematical model to account more fully for the redundancy in the genetic code, and used more robust statistical methods to fit the model to the data. The same cytochrome *c* data was used as in Fitch and Markowitz (1970) and Fitch (1971a). This gave an average number of covarions of at most five, with about 35 to 65% of the covarions losing variability after each substitution. Karon also compared the covarion model with Holmquist, Cantor, and Jukes' random evolutionary hit (REH) interactive model (Holmquist *et al.*, 1972), and found that both models fit the data well, and thus both may be valid.

Fifteen years later, Fitch and Ayala (1994) used computer simulations to show that Cu,Zn superoxide dismutase (SOD), which had earlier been found to behave in an apparently very unclocklike manner (Ayala, 1986), could be a fairly accurate molecular clock under the covarion model, given an appropriate set of parameters. The parameters used were (*i*) sequence length of 118 potentially variable amino acids (out of a total of 162 codons), (*ii*) number of covarions = 28, (*iii*) persistence of variability = 0.01 (persistence of variability

has a different meaning in this paper than in the earlier studies, being the probability that no covarion will be exchanged for a presently invariable codon; only one of the covarions can be exchanged after each substitution—this makes it a somewhat more restrictive parameter), (*iv*) an average of 2.5 alternative amino acids at each variable site, and (*v*) 6 replacements per 10 million years.

Miyamoto and Fitch (1995) performed a more detailed simulation analysis of a subset of the SOD data of Ayala and Fitch, comparing the covarion model with both the Jukes-Cantor one-parameter model and the one-parameter process with a gamma distribution of rates across sites (Nei and Gojobori, 1986; Nei, 1991). The study focused on the difference between the varied and unvaried codons of mammals and plants, and found this to be more consistent with the covarion model than with either of the other models examined.

A few other papers have incorporated aspects of the covarion hypothesis into more general studies (for example Koonin and Gorbalenya, 1989; Fitch and Ye, 1991; Marshall *et al.*, 1994). Most papers that refer to covarions, however, only do so in passing, usually either pointing out possible examples of covarions (e.g. Bogardt *et al.*, 1976; Penny *et al.*, 1987), or simply citing the covarion hypothesis as established fact (e.g. Holmquist, 1972; Penny, 1974; Czelusniak *et al.*, 1978; Golding, 1983; Palumbi, 1989; Dorit and Ayala, 1995). A fairly comprehensive search of the literature only revealed one author (Gillespie, 1986 and 1988) who seems to disagree with the covarion hypothesis, out of over a hundred papers. Gillespie claims that the covarion model is simply an extreme example of a model in which some sites evolve more rapidly than others, apparently ignoring the main point of the covarion model, which is that different sites are free to change at different times.

1.4 The model

The simulations described in this thesis use a simplified covarion-style model. The covarion hypothesis as originally formulated was considered too complex because each site could be 'on' or 'off' (variable or fixed) on different parts of the tree. This appeared

to require a large number of parameters, since sites could only switch between the 'on' and 'off' states as a result of a substitution in one of the 'on' sites. With such a large number of parameters, almost any tree could be made to fit a given data set. One approach suggested as a way around the problem of too many parameters was to use a hidden Markov model (HMM).

Hidden Markov models (Baum and Petrie, 1966) have been applied to a number of problems in molecular biology over the last decade. Lander and Green (1987) used HMMs in the construction of genetic linkage maps, and they have also been used to distinguish coding from non-coding regions in DNA (Churchill, 1989). Simple HMMs have been used in conjunction with the Expectation-Maximisation algorithm to model certain protein-binding sites in DNA (Lawrence and Reilly, 1990; Cardon and Stormo, 1992). Protein families (Krogh *et al.*, 1994; Hughey and Krogh, 1996; Barrett *et al.*, 1997) and superfamilies (Stultz *et al.*, 1993) have been modelled using HMMs. HMMs have also been applied to multiple sequence alignment of proteins (Haussler *et al.*, 1993; Baldi *et al.*, 1994; Krogh *et al.*, 1994; Eddy,1995; Eddy *et al.*, 1995; Hughey and Krogh, 1996; McClure *et al.*, 1996; Barrett *et al.*, 1997), as well as protein structure prediction (Asai *et al,* 1993; Hubbard and Park, 1995). Mitchison and Durbin (1995) developed a tree-based HMM for maximum likelihood evolutionary trees which allows insertions and deletions, and Felsenstein and Churchill (1996) used an HMM to model variation in evolutionary rates among sites.

A hidden Markov model is a Markov model where the states of the system are not directly observable, but the observation is a probabilistic function of the state. That is, the HMM is a doubly embedded stochastic process with an underlying stochastic process that is hidden, but can be observed through another set of stochastic processes that produce the sequence of observations (Rabiner, 1989).

In the case of the covarion hypothesis, the 'hidden' part of the model would be changes between variable and fixed states at each site, and the observations would consist of the character-state (nucleotide or amino acid) at each site. Since this is such a simple example of an HMM, it can actually be reformulated as a single stochastic process.

The process modelled in these simulations (see figure 1) is based on the Kimura 3ST model, but with sites switching between fixed and variable states as well as between the four nucleotides. This gives us a total of eight character states, since each of the four nucleotides can be either fixed or variable at any given time. We will use the notations N_f and N_v (where N can be A, G, C, or T) for fixed and variable nucleotides respectively. This model only requires two additional parameters—the proportion of sites that are variable at any given time (that is, the number of covarions), and the rate of exchange between fixed and variable states. This model is essentially the same as that of Tuffley and Steel (1996), who analysed a simple covarion-style model based on an i.i.d. model but with an additional two-state Markov process that switches sites between the fixed and variable states. They showed that this model cannot be distinguished from one with a static distribution of rates across sites by pairwise comparison of sequences, but that the two models can be distinguished when there are at least four monophyletic groups of taxa.

This thesis explores the effects of varying the rate of exchange between fixed and variable states, as well as the proportion of sites that are covarions. Sequence and distance data are generated by computer simulation on a given tree, and these data are used to test the performance of two tree-building methods (maximum likelihood and neighbour joining) at recovering the original tree.

Even though these tree-building methods assume i.i.d. mechanisms, they are expected to be able to provide a good estimate of the correct tree after longer periods of time with data generated under the covarion-style model than with data generated under i.i.d. models, because more information will be lost due to multiple substitutions under an i.i.d. model. The neighbour joining and maximum likelihood methods tested were in fact found to perform better with data generated under the covarion model, suggesting that current tree-building methods may perform better with biological data than previous computer simulation studies have suggested.

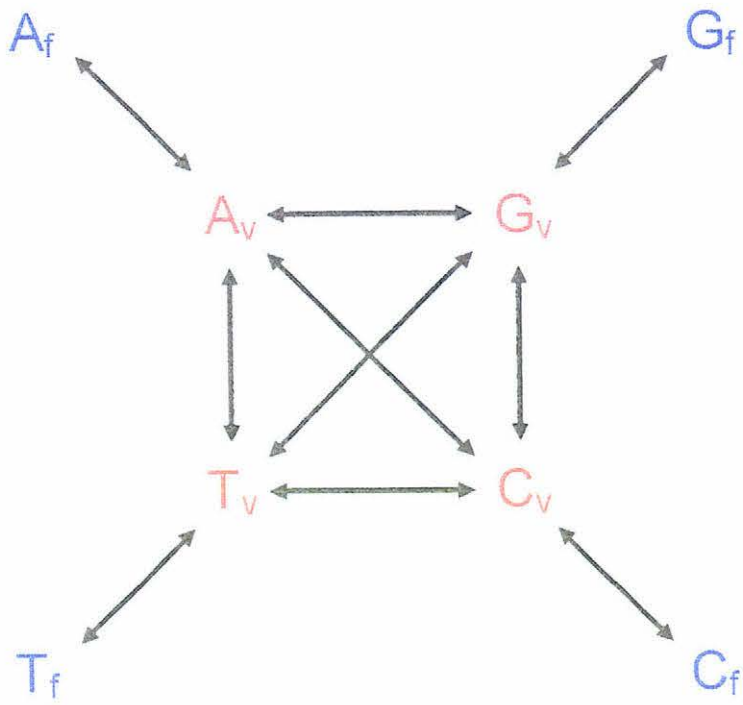


FIGURE 1: THE MODEL

The covarion-style model used in this study. The subscripts f and v denote fixed and variable states respectively.

2. Methods

2.1 The simulations

Simulations were carried out using a program (covarion.exe, written by D. Penny) that allows the user to specify the following parameters: sequence length, c; e, the rate of exchange between fixed and variable states (that is, the expected number of fixed $\leftrightarrow$ variable exchanges per nucleotide substitution); f, the proportion of sites that are permanently fixed; the initial nucleotide composition; the number of simulations, s, to be performed; the proportion of sites that are variable, v; the number of character states, r; number of taxa, n; a tree topology with edge weights; and α, β, and γ values of the Kimura 3ST instantaneous rate matrix, **M**. The output can be in the form of sequences, distances, or both.

All of the simulations used the same basic four character-state 2ST instantaneous rate matrix, with $\alpha = 0.005$ and $\beta = \gamma = 0.0025$, that is,

$$\mathbf{M} = \begin{array}{c} \\ \mathrm{A} \\ \mathrm{G} \\ \mathrm{C} \\ \mathrm{T} \end{array} \begin{array}{c} \begin{array}{cccc} \mathrm{A} & \mathrm{G} & \mathrm{C} & \mathrm{T} \end{array} \\ \begin{bmatrix} -0.01 & 0.005 & 0.0025 & 0.0025 \\ 0.005 & -0.01 & 0.0025 & 0.0025 \\ 0.0025 & 0.0025 & -0.01 & 0.005 \\ 0.0025 & 0.0025 & 0.005 & -0.01 \end{bmatrix} \end{array}.$$

This means that, for example, the rate per unit time interval at which a site moves from state G to state C is 0.0025, and the total rate at which a site leaves state G is 0.01. For a background in continuous-time Markov processes, the reader is referred to Anderson (1991).

From this matrix, along with the rate of exchange, e, and the proportion of sites that are variable, v, the program calculates an eight-by-eight instantaneous rate matrix, **R**. For example, with $e = 1$ and $v = 0.5$,

$$\mathbf{R} = \begin{array}{c} \\ A_v \\ G_v \\ C_v \\ T_v \\ A_f \\ G_f \\ C_f \\ T_f \end{array} \left[\begin{array}{cccc|cccc} -0.02 & 0.005 & 0.0025 & 0.0025 & 0.01 & 0 & 0 & 0 \\ 0.005 & -0.02 & 0.0025 & 0.0025 & 0 & 0.01 & 0 & 0 \\ 0.0025 & 0.0025 & -0.02 & 0.005 & 0 & 0 & 0.01 & 0 \\ 0.0025 & 0.0025 & 0.005 & -0.02 & 0 & 0 & 0 & 0.01 \\ \hline 0.01 & 0 & 0 & 0 & -0.01 & 0 & 0 & 0 \\ 0 & 0.01 & 0 & 0 & 0 & -0.01 & 0 & 0 \\ 0 & 0 & 0.01 & 0 & 0 & 0 & -0.01 & 0 \\ 0 & 0 & 0 & 0.01 & 0 & 0 & 0 & -0.01 \end{array}\right] .$$
$$\begin{array}{cccccccc} A_v & G_v & C_v & T_v & A_f & G_f & C_f & T_f \end{array}$$

The top left-hand quarter of this matrix is essentially the same as the four-by-four matrix above, except for the diagonal elements, which have been altered so that each row sums to zero. A transition matrix, **P**, for each edge is calculated from the eight-by-eight instantaneous rate matrix as follows:

$$\mathbf{P} = e^{t\mathbf{R}},$$

that is, **R** is multiplied by the edge weight t, and the matrix exponential is taken, where the matrix exponential is calculated as follows:

$$e^{\mathbf{A}} = \sum_{n=0}^{\infty} \left(\frac{1}{n!}\right) \mathbf{A}^n$$

For example, the transition matrix generated from the rate matrix above with $t = 50$ is (to five decimal places)

$$\mathbf{P} = \begin{array}{c} \\ A_v \\ G_v \\ C_v \\ T_v \\ A_f \\ G_f \\ C_f \\ T_f \end{array} \begin{array}{c} \begin{array}{cccc|cccc} A_v & G_v & C_v & T_v & A_f & G_f & C_f & T_f \end{array} \\ \left[\begin{array}{cccc|cccc} 0.44406 & 0.10980 & 0.06504 & 0.06504 & 0.25232 & 0.03005 & 0.01685 & 0.01685 \\ 0.10980 & 0.44406 & 0.06504 & 0.06504 & 0.03005 & 0.25232 & 0.01685 & 0.01685 \\ 0.06504 & 0.06504 & 0.44406 & 0.10980 & 0.01685 & 0.01685 & 0.25232 & 0.03005 \\ 0.06504 & 0.06504 & 0.10980 & 0.44406 & 0.01685 & 0.01685 & 0.03005 & 0.25232 \\ \hline 0.25232 & 0.03005 & 0.01685 & 0.01685 & 0.67293 & 0.00527 & 0.00287 & 0.00287 \\ 0.03005 & 0.25232 & 0.01685 & 0.01685 & 0.00527 & 0.67293 & 0.00287 & 0.00287 \\ 0.01685 & 0.01685 & 0.25232 & 0.03005 & 0.00287 & 0.00287 & 0.67293 & 0.00527 \\ 0.01685 & 0.01685 & 0.03005 & 0.25232 & 0.00287 & 0.00287 & 0.00527 & 0.67293 \end{array}\right] \end{array}$$

A value p_{ij} in the matrix represents the probability that a site that is in state i at one end of the edge will be in state j at the other end. For example, the probability of a site in state C_v changing to state T_f on this edge is 0.03005.

The initial nucleotide composition in all the simulations was 0.25:0.25:0.25:0.25, no sites were permanently fixed, and $s = 1000$ simulations were performed for each set of parameters. Unless otherwise specified, $c = 1000$ and $v = 0.5$. The output from each set of simulations was a sequence file and a distance file, both in PHYLIP input file format.

2.2 The trees

Two tree topologies were used in the simulations: the four-taxon tree shown in figure 2(a) and the five-taxon tree in figure 2(b). Although these are shown as rooted trees, the simulations treat them as unrooted. Since the model is time-reversible, the sequences are essentially evolved from one of the final taxa (chosen arbitrarily) rather than from an initial root distribution, so that the position of the root does not need to be specified. The edge weights x and y (see figure 2) were given a range of values between 5 and 690, and z was initially given a value of 5. (An edge weight of 100 corresponds to an average of one substitution per site, or roughly 260 to 450 million years at the neutral substitution rate, estimated by Bulmer *et al.* (1991) as 2.2×10^{-9} to 3.8×10^{-9}.)

Consequently, the internal structure of each tree is fixed while the lengths of the external edges are increased, to simulate increasing time since the taxa diverged. Simulations were also carried out on the five-taxon tree with y fixed at 240 and z ranging from 5 to 80, to test whether information could be recovered from sequences that would normally be randomised under an i.i.d. model due to the length of the external edges.

First, a series of simulations was run on the four-taxon tree, with the rate of exchange, e, ranging between 0.001 and 5. Simulations were then run with $e = 1$ and sequence lengths of 100 and 500, to test whether the effects of increasing rate of exchange between fixed and variable states are similar to those of increasing sequence length. Simulations were run on the five-taxon tree with e ranging between 0.001 and 5. A set of simulations was also run with $v = 1$ (all sites variable) and all rates in the instantaneous rate matrix halved, which is equivalent to setting e at infinity.

The effect of v, the proportion of variable sites, was investigated by running a series of simulations on the five-taxon tree with $e = 1$, $y = 70$ and $z = 5$, with values of v from 0.01 to 1. Since this meant that the overall substitution rate was different for each value of v, another series of simulations was run with the edge weights adjusted such that the expected number of substitutions on the tree was the same (0.8 substitutions per site between the root and any external node) for all values of v. This was done by dividing all edge weights by $2v$, that is, $y = 70/2v$ and $z = 5/2v$.

2.3 Tree-building

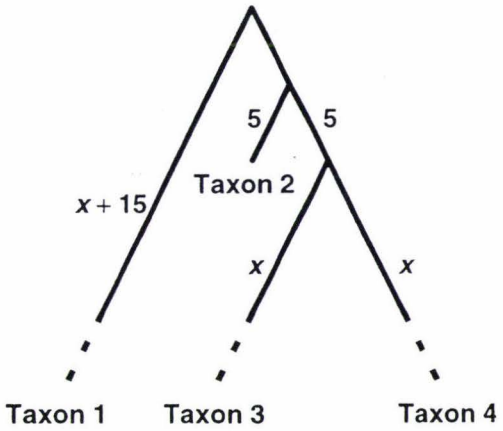
Trees were reconstructed from the simulated data (the output from covarion.exe) using programs from PHYLIP version 3.572 (Felsenstein, 1993). DNAML and DNAPARS were used for maximum likelihood and parsimony, respectively, on the sequence data, and NEIGHBOR for neighbour-joining on the distances.

Due to the computational time involved (especially for the maximum likelihood tree-building, which in some cases can take 48 hours or more to run on a 133MHz Pentium for a set of a thousand five-taxon trees), the simulation and tree-building steps were

performed using up to 51 Pentiums running in parallel in a student computing laboratory, controlled remotely from a network account. This process was largely automated by using a batch file and input files to control each set of simulations and transfer its output into the tree-building programs. One of these batch files is given in the appendix, along with the associated input files.

The output from the tree-building programs was in the form of tree files of a thousand trees. The output files for the four-taxon tree were analysed using a program (nexustr4.exe, written by D. Penny) that counts the number of times each possible tree occurs and gives the proportion of simulations for which the correct tree is inferred. For the five-taxon tree files, a similar program (nexustre.exe, written by D. Penny) was used, that gives the proportion of simulations for which 0, 1, or 2 edges are wrongly inferred. An example of an output file from each of these programs is given in the appendix.

(a)



(b)

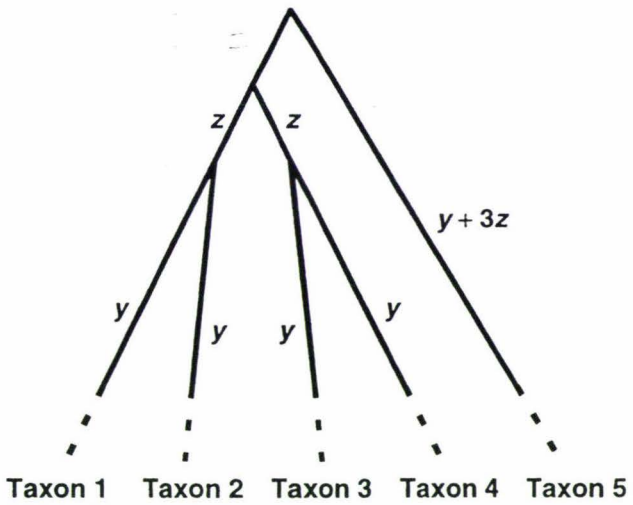


FIGURE 2: THE TREES

The two basic trees used in the simulations: (a) a four-taxon tree in which the weights of the internal edge and one of the four pendent edges are fixed; and (b) a five-taxon tree in which the weights of both internal edges are fixed. The edge weight parameters x and y take a range of values to simulate increasing time since the initial branching.

3. Results

3.1 Four-taxon simulations

Both of the tree-building programs tested performed better (that is, they had a higher probability of selecting the correct tree) with data from simulations that used higher values of e, the rate of exchange between fixed and variable states. This is illustrated in figure 3. The effect of increasing e was similar to that of increasing the length of the sequences, as shown in figure 4. This implies that the effect of increasing the rate of exchange between fixed and variable states is in some way equivalent to an increase in 'effective sequence length'.

Even though the same data was used for both programs, DNAML performed better than NEIGHBOR for all values of e tested. As table 1 shows, increasing e from 0.001 to 5 improves the performance of both programs first at low confidence levels and later at the higher confidence levels. For example, increasing e from 0.001 to 0.5 allows a 134% increase in the tree length for which DNAML is 50% correct but only a 10% increase in tree length at the 95% confidence level, while increasing e from 0.5 to 1.5 allows a 32% increase in tree length at the 95% confidence level but no increase in tree length at 50% confidence.

An interesting anomaly can be seen in figure 3, in that, for low values of e, the lines drop below the percentage of correct trees expected by chance (33.3%), before levelling off at this value. In order to check whether this was significant, 18 sets of 1000 simulations were carried out with $e = 0.001$ and $x = 190$. The mean percentage of correct trees was found to be 29.3% for DNAML and 28.9% for NEIGHBOR, with standard deviations of 1.5 and 1.4 respectively. Both tree-building programs found the correct tree less than 33.3% of the time for all 18 replicates. The anomaly therefore appears to be significant. This could be due to a long edge attraction effect (Hendy and Penny, 1989) causing the longest external edge (leading to taxon 1) to pair with either of the other long edges more

often than it pairs with the short edge leading to taxon 2, because of parallel changes accumulating on the long edges. Further testing would be required to confirm this.

3.2 Five-taxon simulations

The effects of increasing e (the rate of exchange) were found to be the similar for five taxa as for four, in that the performance of the tree-building programs improved as e increased (see figure 5). However, DNAML only performed better than NEIGHBOR for e greater than 1. As table 2 shows, NEIGHBOR performs better than DNAML for low values of e. Even at the highest values of e, DNAML performs only slightly better, whereas for the four-taxon tree DNAML did substantially better than NEIGHBOR for all values of e.

When the external edges of the tree were fixed ($y = 240$) and the internal structure was lengthened (z ranging from 5 to 80), the tree was recovered most successfully with relatively low values of e (0.1 with DNAML and 0.5 with NEIGHBOR; see figure 6). The reason for this is probably that when e is high, more sites are free to accept substitutions at some point on each of the long external edges, so that the sequences are closer to being randomised. Conversely, if e is too low, any sites that fix substitutions on the internal edges remain free to vary on the external edges, so that any useful information that may have been generated is lost as those sites are randomised. For this set of simulations, NEIGHBOR performed better than DNAML in all cases. This is surprising, since maximum likelihood is generally more robust than neighbour joining against differences in the model of evolution, such as the presence of invariant sites (Lockhart *et.al.*, 1996). Serious study is required to clarify the behaviour of the different tree-building methods under the covarion model.

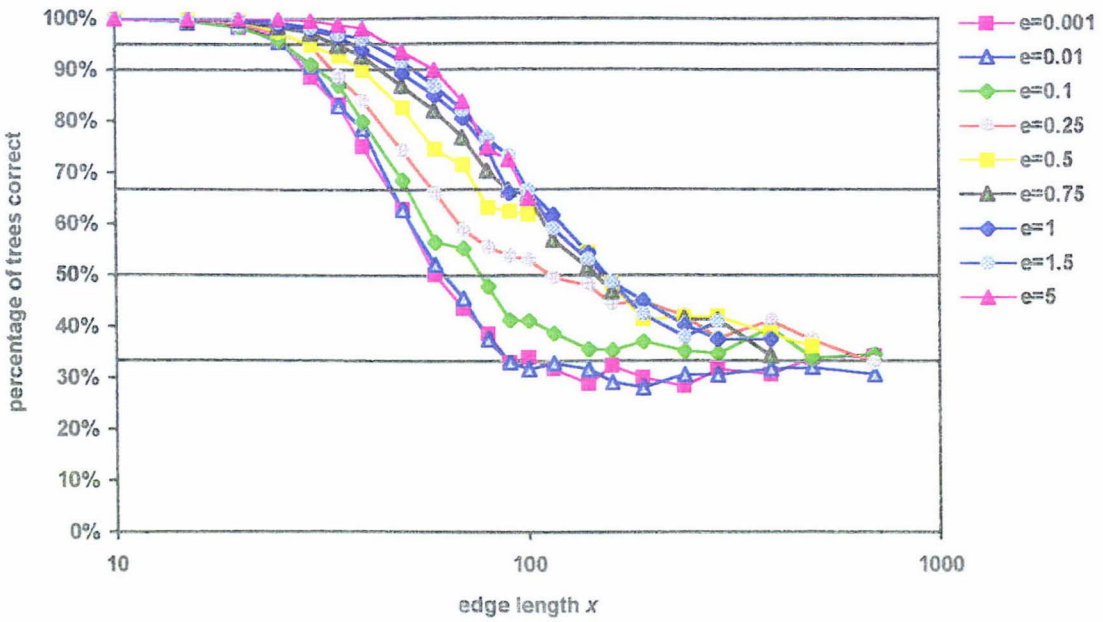
All of the results above were for simulations with v, the proportion of variable sites, fixed at 0.5. Changing v gave the curve shown in figure 7, with the tree-building programs performing best with v around 0.5. However, this effect could be due to the fact that with the unadjusted edge weights, the overall substitution rate is proportional to v (since the rate per variable site is constant), so that for low values of v the number of substitutions

is too small to provide enough information to reconstruct the tree reliably, while for high values of v the information is lost due to multiple substitutions. When the edge weights were adjusted to give the same overall rate despite the different proportions of variable sites, this effect disappeared and changing v had little or no effect on the programs' performance. The proportion of sites that are variable therefore does not have a large effect on the performance of the tree-building programs used, as long as the rate of substitution per site (a parameter that is relatively easy to estimate from real data) is fixed.

FIGURE 3: PERFORMANCE OF TREE-BUILDING PROGRAMS ON THE FOUR-TAXON TREE WITH DIFFERENT RATES OF EXCHANGE

The percentage of trials in which (a) DNAML and (b) NEIGHBOR infer the correct four-taxon tree is shown for values of e ranging from 0.001 to 5, and for values of x (see figure 2) up to 690. The edge lengths x are plotted on a log scale. The horizontal line at 33.3% shows the percentage of trials in which the correct tree should be inferred by chance from random sequences.

(a) DNAML



(b) NEIGHBOR

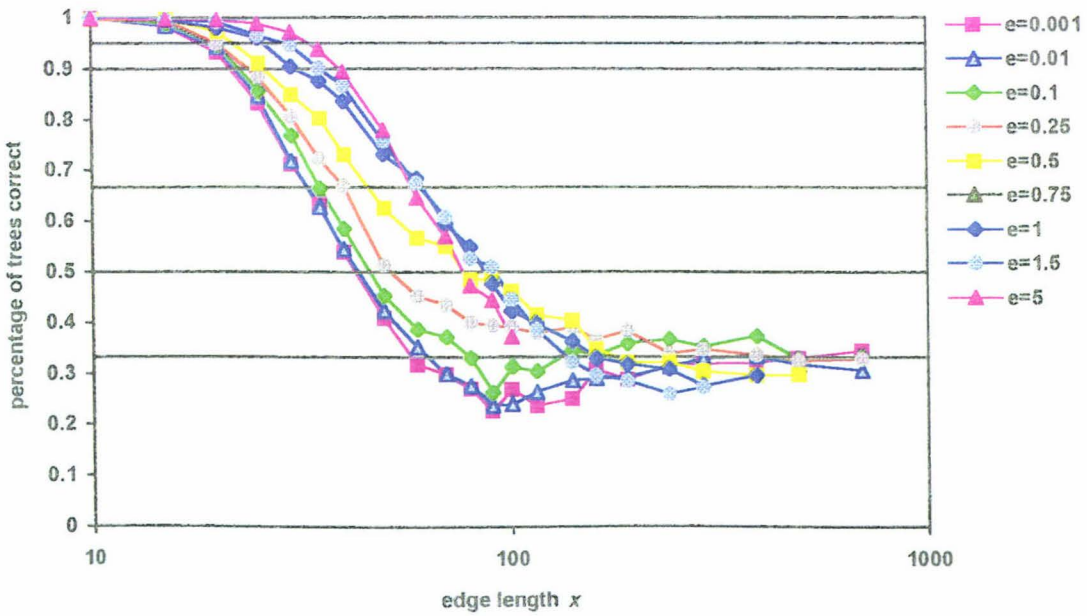


TABLE 1: VALUES OF x FOR WHICH THE CORRECT FOUR-TAXON TREE IS INFERRED WITH 50, 67, 90, AND 95% PROBABILITY

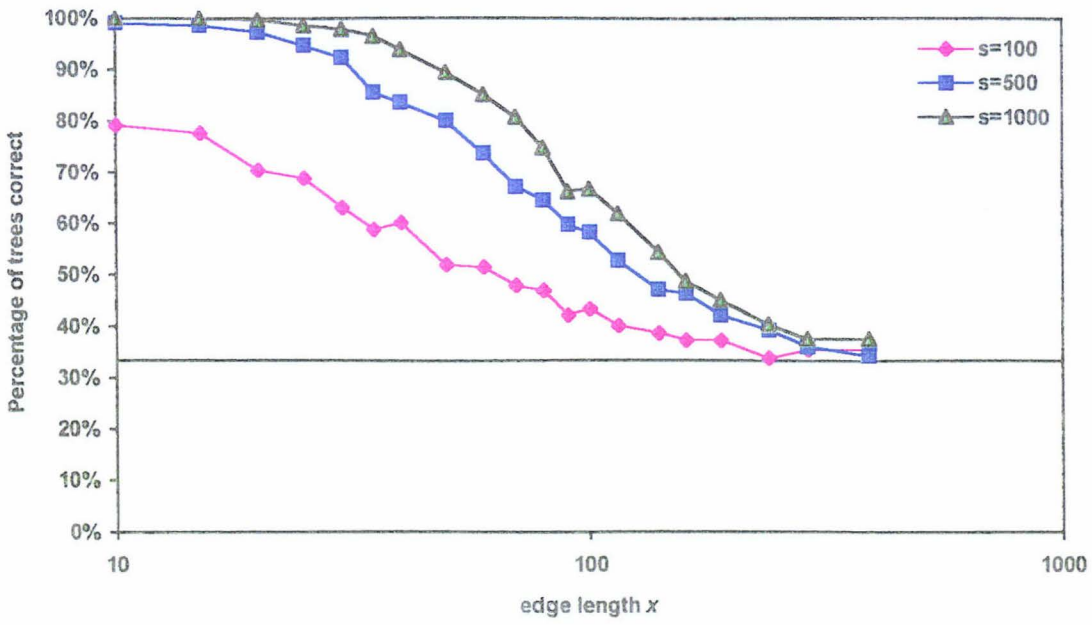
Values of the edge length x for which DNAML and NEIGHBOR infer the correct tree in 50%, 67%, 90% and 95% of trials, with different values of the rate of exchange, e, between fixed and variable sites. Values of x were calculated by linear interpolation from the two values between which the lines in figure 3 cross each probability value (shown as horizontal lines in figure 3).

	DNAML				NEIGHBOR			
e	50%	67%	90%	95%	50%	67%	90%	95%
0.001	60.5	47.1	29.2	26.0	43.0	32.9	21.7	18.4
0.01	63.5	47.6	30.6	25.6	43.8	32.9	22.3	19.3
0.1	77.3	51.7	31.4	25.9	46.5	34.9	22.6	19.7
0.25	113.7	59.4	34.0	29.8	52.2	40.2	23.9	20.0
0.5	155.2	75.9	40.0	29.8	77.8	46.2	26.0	21.9
0.75	147.1	90.3	44.7	34.6	83.1	54.8	29.6	24.0
1	155.8	102.1	48.7	38.0	86.8	62.1	30.9	26.1
1.5	154.2	99.9	53.8	42.6	91.7	61.3	35.3	29.3
5	—	97.9	60.2	47.0	77.3	58.5	39.5	33.5

FIGURE 4: PERFORMANCE OF TREE-BUILDING METHODS WITH DIFFERENT SEQUENCE LENGTHS

The percentage of trials in which (a) DNAML and (b) NEIGHBOR infer the correct four-taxon tree is shown for sequence lengths s = 100, 500 and 1000, and for values of x (see figure 2) up to 690. The edge lengths x are plotted on a log scale. The horizontal line at 33.3% shows the percentage of trials in which the correct tree should be inferred by chance from random sequences.

(a) DNAML



(b) NEIGHBOR

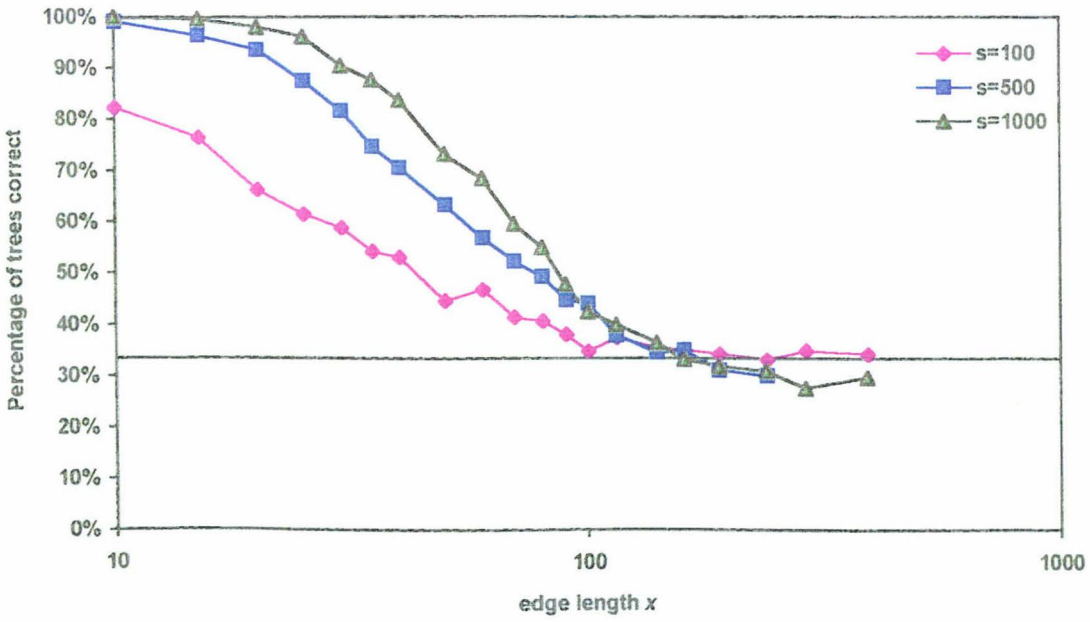
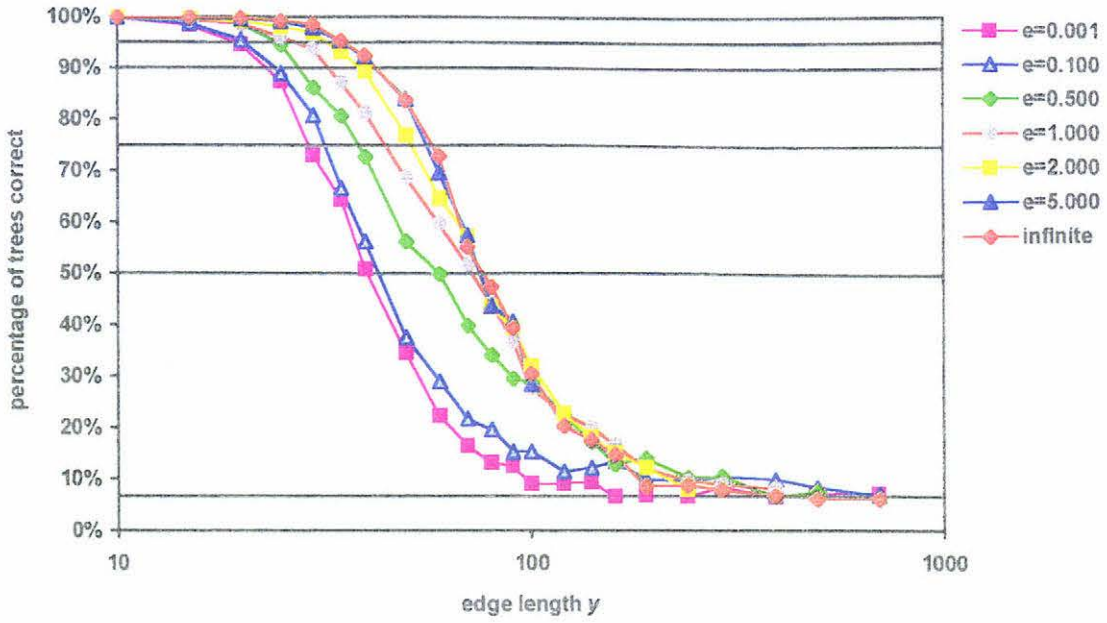


FIGURE 5: PERFORMANCE OF TREE-BUILDING PROGRAMS ON THE FIVE-TAXON TREE WITH DIFFERENT RATES OF EXCHANGE, PART 1

The percentage of trials in which (a) DNAML and (b) NEIGHBOR infer the correct five-taxon tree is shown for values of e ranging from 0.001 to infinity, and for values of y (see figure 2) up to 690. In order to give e a value of infinity, the proportion of variable sites, v, was set to 1 and all values in the instantaneous rate matrix **M** were halved. The edge lengths y are plotted on a log scale. The horizontal line at 6.67% shows the percentage of trials in which the correct tree should be inferred by chance from random sequences.

(a) DNAML



(b) NEIGHBOR

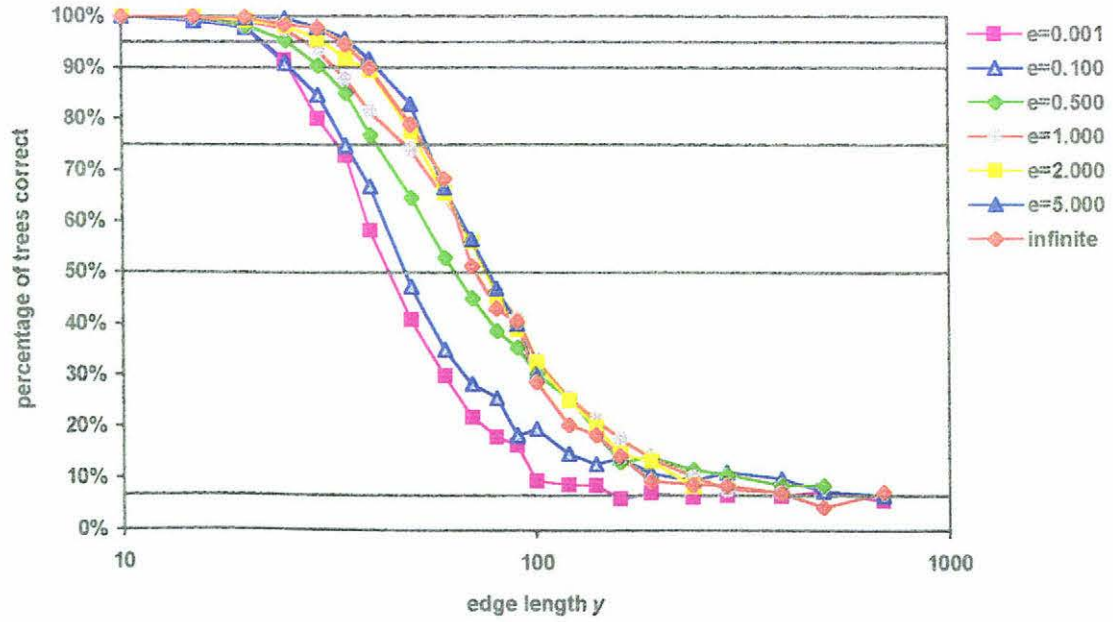


TABLE 2: VALUES OF *Y* FOR WHICH THE CORRECT FIVE-TAXON TREE IS INFERRED WITH 50, 67, 90, AND 95% PROBABILITY

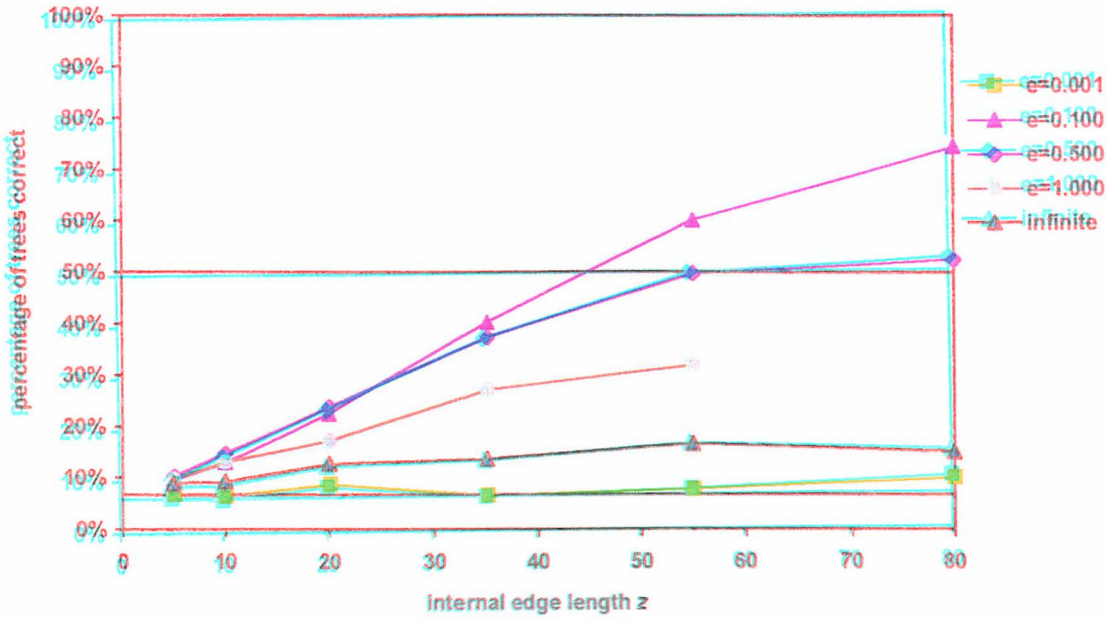
Values of the edge length y for which DNAML and NEIGHBOR infer the correct tree in 50%, 67%, 90% and 95% of trials, with different values of the rate of exchange, e, between fixed and variable sites. Values of y were calculated by linear interpolation from the two values between which the lines in figure 5 cross each probability value (shown as horizontal lines in figure 5).

	DNAML				NEIGHBOR			
e	50%	75%	90%	95%	50%	75%	90%	95%
0.001	40.6	29.3	23.2	19.6	44.8	33.5	25.7	22.3
0.1	43.4	32.1	24.3	20.5	48.6	34.9	25.8	22.1
0.5	60.0	38.6	27.7	24.5	63.8	41.6	30.5	25.3
1	72.2	45.1	32.8	27.3	76.2	48.8	33.1	28.2
2	75.3	51.6	39.1	32.5	75.7	52.1	39.0	30.8
5	75.5	56.3	42.8	35.6	76.9	54.9	42.0	35.9
∞	76.8	58.2	42.9	35.7	71.7	53.8	40.1	34.5

FIGURE 6: PERFORMANCE OF TREE-BUILDING PROGRAMS ON THE FIVE-TAXON TREE WITH DIFFERENT RATES OF EXCHANGE, PART 2

The percentage of trials in which (a) DNAML and (b) NEIGHBOR infer the correct five-taxon tree with external edge weights y fixed at 240 is shown for values of e ranging from 0.001 to infinity, and for values of z (see figure 2) up to 80. In order to give e a value of infinity, the proportion of variable sites, v, was set to 1 and all values in the instantaneous rate matrix **M** were halved. The horizontal line at 6.67% shows the percentage of trials in which the correct tree should be inferred by chance from random sequences.

(a) DNAML



(b) NEIGHBOR

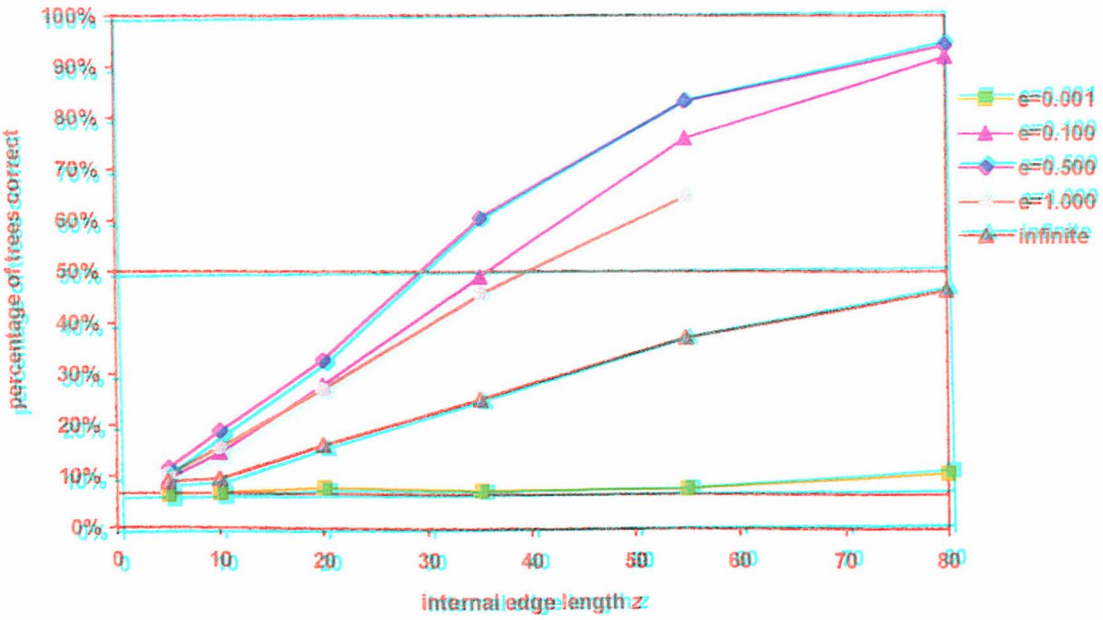
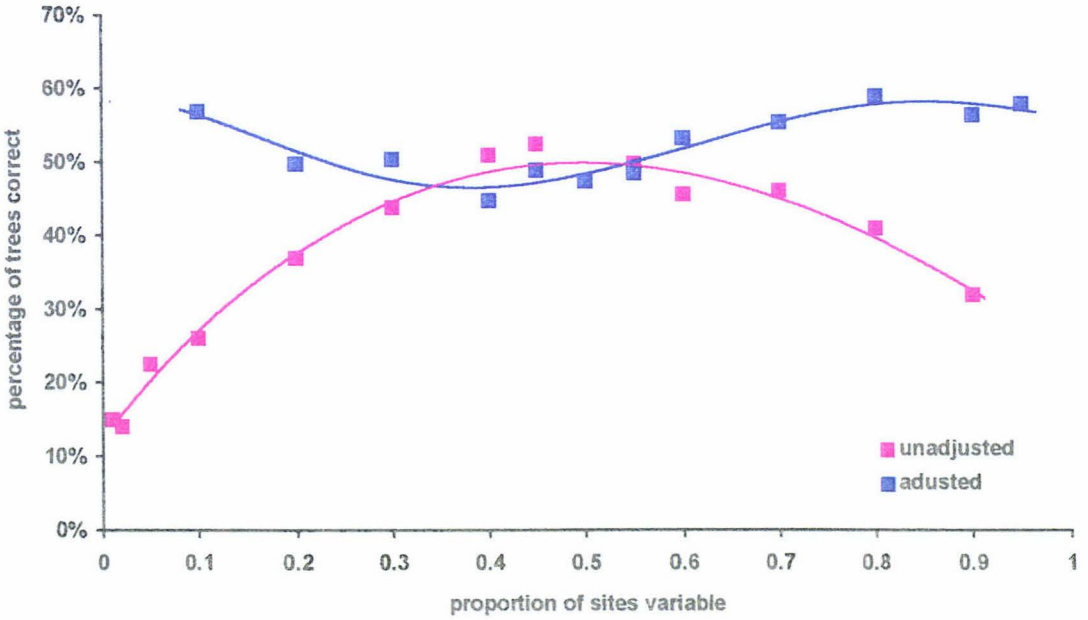


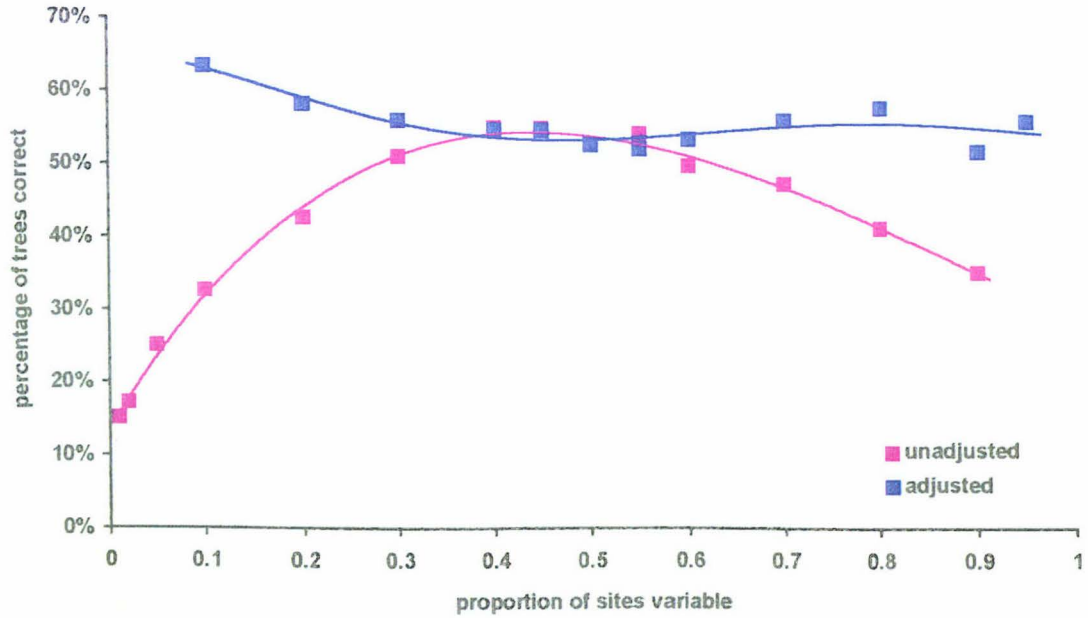
FIGURE 7: PERFORMANCE OF TREE-BUILDING PROGRAMS ON THE FIVE-TAXON TREE WITH DIFFERENT PROPORTIONS OF VARIABLE SITES

The percentage of trials in which (a) DNAML and (b) NEIGHBOR infer the correct five-taxon tree is shown for values of v ranging from 0.01 to 1. For the lines labelled 'unadjusted,' the edge lengths are $y = 70$ and $z = 5$, while for the lines labelled 'adjusted', the edge lengths are inversely proportional to v, that is, $y = 70/2v$ and $z = 5/2v$, so that the overall number of substitutions per site is fixed at 0.4.

(a) DNAML



(b) NEIGHBOR



4. Discussion

The main outcomes of this study are to show that the covarion hypothesis can be successfully modelled as a simple hidden Markov model, and that relationships generated using this model can be inferred further back in time than those generated with i.i.d. models, even when the models used to reconstruct the trees are i.i.d. This means that if the covarion hypothesis describes a realistic biochemical mechanism of evolution (and it seems to be generally accepted that it does), and if the model described in this study is a good approximation of the covarion hypothesis, then the tree-building programs tested should perform better with real sequence data than they do with data simulated under an i.i.d. model. It seems likely, therefore, that somewhat older divergences can be reconstructed reliably from sequences than is currently thought possible, even with current tree-building methods that assume an i.i.d. mechanism of change. For some of the conditions modelled here, divergences about 50 to 75% older can be reconstructed than under an i.i.d. model, but there may be conditions under which the covarion model could perform even better. The model parameters need to be explored further to find such conditions, and the parameters should be estimated from real data to see whether the optimal conditions are realistic.

The results of Miyamoto and Fitch (1995) are consistent with a gamma distribution of rates across sites being a closer approximation to the covarion model than the one-parameter i.i.d. model, although observed superoxide dismutase (SOD) data fit the covarion model better than the gamma model because different sites were found to have changed in plants and mammals, so that the rate at a given site is not constant across the tree. The conclusion that the gamma distribution is a good approximation for the covarion model, is supported by the mathematical analysis of Tuffley and Steel (1996), who obtained a distance measure that is tree-like for certain parameters of a covarion model, but is generally not tree-like under rates-across-sites models. Although this measure can in principle be used to distinguish between the covarion model and rates-across-sites when there are at least four monophyletic groups of taxa, it is not clear

whether the two models can be distinguished on the basis of pairwise comparison of sequences. Since it is difficult to distinguish between the covarion model and a gamma distribution of rates across sites, tree-building programs that incorporate a gamma model may perform better than those using an i.i.d. model, while requiring one less parameter than even the simplified covarion model presented here.

Covarion parameters need to be estimated from a variety of real data sets, to determine realistic parameter values and to ascertain whether the parameters are the same for different genes, or whether they vary. The proportion of sites that are free to change at any point in time is likely to differ greatly among genes due to the diverse constraints on different genes, and it could even vary over time for a single gene. This may also be true for the proportion of sites that are considered to be permanently fixed, especially if there is any change in the gene's function, as noted by Lockhart *et al.* (1996). On the other hand, e, the rate of exchange between fixed and variable sites, may be fairly constant, since the mechanism of exchange should be largely the same for all genes. It would be interesting to test how widely these parameters vary among different genes.

Some studies estimating covarion parameters have been carried out already (those of Fitch (1971a), Karon (1979), and Fitch and Ayala (1994); see introduction for details), but these used slightly different parameters than those in the present model, using 'persistence of variability' rather than the rate of exchange between fixed and variable sites. Persistence of variability is a more restrictive parameter than e, since it assumes that an exchange between fixed and variable sites can only occur after a substitution, and that such exchanges are strictly on a one-for-one basis, that is, one covarion can be exchanged for one fixed site after each substitution. This assumption is unrealistic from a biological point of view (as Miyamoto and Fitch (1995) recognised), since changes in external constraints on the molecule (such as mutations in other molecules that interact physically, or changes in temperature or pH) could result in changes in the covarion set, or a single mutation could cause several covarions to be exchanged for fixed sites.

In order to estimate covarion parameters for a gene, it is necessary to have a phylogeny for the sequences that is based on data independent of the gene in question and that is

widely accepted. Palaeontological dates need to be estimated for some divergences in the phylogeny. Ideally, the sequences should come from a large and diverse group of taxa (for example, Fitch and Ayala used 67 SOD sequences from plants, animals, fungi, bacteria, and even a virus). Of course, the larger the number of species, the less likely it is that a robust, independently supported tree with dates will be available. Given such a tree, it is relatively easy to estimate the proportions of sites that are variable and that are permanently fixed, as shown by the studies cited above. The rate of exchange between fixed and variable sites, and the underlying substitution rate, however, are likely to be somewhat more difficult to determine, but could probably be done using some form of maximum likelihood analysis.

Extreme values of some parameters need to be explored further, as the current simulation program does not allow for values greater than or equal to 10 in the eight-by-eight instantaneous rate matrix. This study only considered the topologies of the reconstructed trees, but it would also be interesting to examine how their edge lengths relate to those used in the simulations. This would shed some light on the workings of the molecular clock, which can appear very unreliable if the covarion model is not taken into account, as shown by Fitch and Ayala (1994).

Appendix

This appendix gives an example of each of the files involved in running a set of simulations remotely over the local Massey network, as well as output files from the programs nexustre.exe and nexustr4.exe.

A.1 Batch file

A batch file (100e2000.bat) used for a set of simulations and tree-building. Comments are given to explain each step in the file.

```
c:
cd \user\sims
covarion < h:\100e2000.sim
```

{Executes the simulation program covarion.exe and reads in the input file 100e2000.sim.}

```
ren covarion.log 100e2000.log
```

{Renames the file covarion.log to 100e2000.log.}

```
neighbor < h:\dis.in
```

{Executes the tree-building program NEIGHBOR and reads in the input file dis.in.}

```
ren treefile 100e2000.nj
```

{Renames the file treefile (the output from NEIGHBOR) to 100e2000.nj.}

```
del covarion.dis
```

{Deletes the file covarion.dis (to save space on the hard disk).}

```
dnaml < h:\seq.in
```

{Executes the tree-building program DNAML and reads in the input file seq.in.}

```
ren treefile 100e2000.ml
```

{Renames the file treefile (the output from DNAML) to 100e2000.ml.}

```
del covarion.seq
```

{Deletes the file covarion.seq (to save space on the hard disk—each of these files is several megabytes in size).}

```
h:
move c:\user\sims\100e1000.* h:\
```

{Moves the files 100e2000.log, 100e2000.nj and 100e2000.ml to the network drive.}

A.2 Input files

The input file 100e2000.sim for covarion.exe, with comments:

```
d
3
```
{Sets the output format to both sequences and distances.}
```
s
1000
```
{Sets the number of simulations to 1000.}
```
e
2
```
{Sets the rate of exchange between fixed and variable states to 2.}
```
y
```
{Accepts the above values and the default values of the other parameters.}
```
m
h:\5k100mdl.txt
```
{Reads in the model file 5k100mdl.txt.}

The input file dis.in for NEIGHBOR:

```
covarion.dis
```
{Gives the name of the data file (covarion.dis) to read in.}
```
l
```
{Switches on the lower-triangular data matrix option.}
```
2
```
{Suppresses indications of the progress of the run.}
```
m
1000
```
{Allows multiple data sets to be processed, and sets the number of these to 1000.}
```
y
```
{Accepts the options set above, and begins the tree-building.}

The input file seq.in for DNAML and DNAPARS:

```
covarion.seq
```

{Gives the name of the data file (covarion.seq) to read in.}

```
i
```

{Switches off the interleaved input sequences option.}

```
2
```

{Suppresses indications of the progress of the run.}

```
m
1000
```

{Allows multiple data sets to be processed, and sets the number of these to 1000.}

```
y
```

{Accepts the options set above, and begins the tree-building.}

A.3 Model file

The file 5k100mdl.txt, containing the tree model (topology, edge lengths, and rate matrix) to be used by covarion.exe:

```
5 1 1
```

{Specifies the number of taxa (5), the format for the topology (1 = Fitch format), and the number of instantaneous rate matrices (one matrix for the whole tree—it is possible to specify a different matrix for each edge).}

```
  taxon_1     taxon_2     taxon_3     taxon_4     taxon_5
```

{Gives the names of the taxa.}

```
    7
    1    2    6    5    7    3    4
```

{Specifies the number of edges in the tree (7 for a five-taxon tree), and the topology in Fitch format (a format used in the Fitch parsimony algorithm (Fitch, 1971b)).}

```
   90    90    5    105    5    90    90
```

{Gives the edge weights.}

```
  -0.0100   0.0050   0.0025   0.0025
   0.0050  -0.0100   0.0025   0.0025
   0.0025   0.0025  -0.0100   0.0050
   0.0025   0.0025   0.0050  -0.0100
```

{Gives the Kimura 3ST instantaneous rate matrix.}

A.4 Output from nexustr4.exe and nexustre.exe

The output file e0500ml.tre generated by nexustr4.exe. The input files were tree files from DNAML for the four-taxon tree with the rate of exchange between fixed and variable states set to 1.5, and the edge weight x ranging from 10 to 490. Values of x greater than 490 caused an error in the simulation program, and were not used.

```
   file is: 020e0500.ml
```
{The name of the input tree file.}

```
   final   1000      0      0
```
{The number of times each of the three possible tree topologies (The correct tree with taxa 1 and 2 together, then the tree joining taxa 1 and 3, and that joining taxa 1 and 4) is found in the tree file.}

```
   file is: 025e0500.ml
   final    999      1      0
   file is: 030e0500.ml
   final    997      3      0
   file is: 035e0500.ml
   final    975     15     10
   file is: 040e0500.ml
   final    949     27     24
   file is: 045e0500.ml
   final    928     36     36
   file is: 050e0500.ml
   final    900     51     49
   file is: 060e0500.ml
   final    827     82     91
   file is: 070e0500.ml
   final    746    127    127
   file is: 080e0500.ml
   final    716    135    149
   file is: 090e0500.ml
   final    633    180    187
   file is: 100e0500.ml
   final    626    172    202
   file is: 110e0500.ml
   final    619    192    189
   file is: 150e0500.ml
   final    544    217    239
   file is: 170e0500.ml
   final    486    265    249
   file is: 200e0500.ml
   final    415    298    287
   file is: 250e0500.ml
   final    420    289    291
   file is: 300e0500.ml
   final    421    299    280
   file is: 400e0500.ml
   final    387    307    306
   file is: 500e0500.ml
   final    362    304    334
```

The output file e2000nj.tre generated by nexustre.exe. The input files were tree files from NEIGHBOR for the five-taxon tree with the rate of exchange between fixed and variable states set to 2, and the edge weight y ranging from 10 to 240. Values of x greater than 240 caused an error in the simulation program, and were not used.

```
  file is: 020e2000.nj
bipartitions and their frequencies:
     31000     121000
 distribution of edges wrong
   0   1   2
1000   0   0
 mean # of edges wrong    0.00        0.     1000.

  file is: 025e2000.nj
 bipartitions and their frequencies:
     31000      7   1     12 999
 distribution of edges wrong
   0   1   2
999   1   0
 mean # of edges wrong    0.00        1.     1000.

  file is: 030e2000.nj
```
{The name of the input tree file.}

```
 bipartitions and their frequencies:
     3 997     11   5     12 995     13   2     14   1
```
{The bipartitions (see Hendy and Penny (1993) for an explanation of bipartitions) found in the input tree file and their frequencies (the correct tree contains bipartitions 3 and 12).}

```
 distribution of edges wrong
   0   1   2
 992   8   0
```
{The number of trees with 0, 1, and 2 incorrect edges (only the two internal edges can be incorrect, the external edges are always correct).}

```
  mean # of edges wrong    0.01        8.     1000.
```
{The mean number of incorrect edges per tree, the total number of incorrect edges in the tree file, and the total number of trees.}

```
  file is: 035e2000.nj
bipartitions and their frequencies:
     3 989      7   2     11   5     12 993     13   5
    14   6
distribution of edges wrong
  0   1   2
982  18   0
 mean # of edges wrong    0.02        18.     1000.

 file is: 040e2000.nj
bipartitions and their frequencies:
     3 981      7  11     11  14     12 975     13  10
    14   9
distribution of edges wrong
  0   1   2
956  44   0
 mean # of edges wrong    0.04        44.     1000.

 file is: 045e2000.nj
```

```
bipartitions and their frequencies:
     3 962       7  23      10   1      11  20      12 956
    13  20      14  18
distribution of edges wrong
  0   1   2
919  80   1
 mean # of edges wrong     0.08        82.      1000.

 file is: 050e2000.nj
bipartitions and their frequencies:
     3 950       7  29      10   1      11  27      12 944
    13  28      14  21
distribution of edges wrong
  0   1   2
895 104   1
 mean # of edges wrong     0.11       106.      1000.

 file is: 060e2000.nj
bipartitions and their frequencies:
     3 887       5   1       6   3       7  60       9   2
    10   2      11  55      12 882      13  37      14  71
distribution of edges wrong
  0   1   2
776 217   7
 mean # of edges wrong     0.23       231.      1000.

 file is: 070e2000.nj
bipartitions and their frequencies:
     3 812       5  16       6  10       7  91       9  10
    10  14      11  87      12 799      13  86      14  75
distribution of edges wrong
  0   1   2
655 301  44
 mean # of edges wrong     0.39       389.      1000.

 file is: 080e2000.nj
bipartitions and their frequencies:
     3 768       5  23       6  14       7 124       9  19
    10  20      11 116      12 725      13 102      14  89
distribution of edges wrong
  0   1   2
561 371  68
 mean # of edges wrong     0.51       507.      1000.

 file is: 090e2000.nj
bipartitions and their frequencies:
     3 646       5  36       6  30       7 123       9  31
    10  39      11 122      12 689      13 136      14 148
distribution of edges wrong
  0   1   2
454 427 119
 mean # of edges wrong     0.67       665.      1000.

 file is: 100e2000.nj
bipartitions and their frequencies:
     3 603       5  49       6  53       7 134       9  50
    10  50      11 154      12 619      13 146      14 142
distribution of edges wrong
  0   1   2
390 442 168
 mean # of edges wrong     0.78       778.      1000.

 file is: 110e2000.nj
bipartitions and their frequencies:
     3 564       5  65       6  69       7 182       9  58
```

```
    10  62     11 166     12 533     13 146     14 155
distribution of edges wrong
  0   1   2
325 447 228
 mean # of edges wrong    0.90      903.     1000.

 file is: 130e2000.nj
bipartitions and their frequencies:
     3 450      5  93      6 109      7 176      9 108
    10  86     11 179     12 456     13 159     14 184
distribution of edges wrong
  0   1   2
252 402 346
 mean # of edges wrong    1.09     1094.     1000.

 file is: 150e2000.nj
bipartitions and their frequencies:
     3 388      5 118      6 119      7 192      9 124
    10 115     11 186     12 399     13 169     14 190
distribution of edges wrong
  0   1   2
200 387 413
 mean # of edges wrong    1.21     1213.     1000.

 file is: 170e2000.nj
bipartitions and their frequencies:
     3 337      5 127      6 138      7 194      9 131
    10 125     11 170     12 355     13 210     14 213
distribution of edges wrong
  0   1   2
148 396 456
 mean # of edges wrong    1.31     1308.     1000.

 file is: 200e2000.nj
bipartitions and their frequencies:
     3 309      5 162      6 139      7 196      9 144
    10 156     11 187     12 312     13 203     14 192
distribution of edges wrong
  0   1   2
134 353 513
 mean # of edges wrong    1.38     1379.     1000.

 file is: 250e2000.nj
bipartitions and their frequencies:
     3 231      5 197      6 171      7 197      9 160
    10 176     11 199     12 254     13 203     14 212
distribution of edges wrong
  0   1   2
 85 315 600
 mean # of edges wrong    1.51     1515.     1000.
```

References

ANDERSON, W.J. 1991. *Continuous-time Markov chains: an applications-oriented approach*. Springer-Verlag, New York.

ASAI, K., S. HAYAMIZU, and K. HANDA. 1993. Prediction of protein secondary structure by the hidden Markov model. *CABIOS* 9, 141-146.

AYALA, F.J. 1986. On the virtues and pitfalls of the molecular evolutionary clock. *J. Hered.* **77**, 226-235.

BALDI, P., Y. CHAUVIN, T. HUNKAPILLER, and M.A. MCCLURE. 1994. Hidden Markov models of biological primary sequence information. *Proc. Natl. Acad. Sci. USA* **91**, 1059-1063.

BARRETT, C., R. HUGHEY, and K. KARPLUS. 1997. Scoring hidden Markov models. *CABIOS* **13**, 191-199.

BAUM, L.E. and T. PETRIE. 1966. Statistical inference for probabilistic functions of finite state Markov chains. *Ann. Math. Stat.* **37**, 1554-1563.

BOGARDT, R.A., F.E. DWULET, L.D. LEHMAN, B.N. JONES, and F.R.N. GURD. 1976. Complete primary structure of the major component myoglobin of California gray whale (*Eschrichtius gibbosus*). *Biochem.* **15**, 2597-2602.

BULMER, M., K.H. WOLFE, and P.M. SHARP. 1994. Synonymous nucleotide substitution rates in mammalian genes: implications for the molecular clock and the relationship of mammalian orders. *Proc. Natl. Acad. Sci. USA* **88**, 5974-5978.

CARDON, L.R. and G.D. STORMO. 1992. Expectation maximization algorithm for identifying protein-binding sites with variable lengths from unaligned DNA fragments. *J. Mol. Biol.* **223**, 159-170.

CHARLESTON, M.A. 1994. *Factors affecting the performance of phylogenetic methods.* Ph.D. thesis, Massey University.

CHURCHILL, G.A. 1989. Stochastic models for heterogeneous DNA sequences. *Bull. Math. Biol.* **51**, 79-94.

CZELUSNIAK, J., M. GOODMAN, and G.W. MOORE. 1978. On investigating the statistical properties of the populous path algorithm by computer simulation: counterconclusions to those of Tateno and Nei. *J. Mol. Evol.* **11**, 75-85.

DORIT, R.L. and F.J. AYALA. 1995. ADH evolution and the phylogenetic footprint. *J. Mol. Evol.* **40**, 658-662.

EDDY, S. 1995. Multiple alignment using hidden Markov models. In *ISMB-95*, pp. 114-120. AAAI/MIT Press, Menlo Park, California.

EDDY, S.R., G. MITCHISON, and R. DURBIN. 1995. Maximum discrimination hidden Markov models of sequence consensus. *J. Comput. Biol.* **2**, 9-23.

FELSENSTEIN, J. 1993. *PHYLIP (Phylogeny Inference Package) version 3.572.* Distributed by the author, Department of Genetics, University of Washington, Seattle, Washington.

FELSENSTEIN, J. and G.A. CHURCHILL. 1996. A hidden Markov model approach to variation among sites in rate of evolution. *Mol. Biol. Evol.* **13**, 93-104.

FITCH, W.M. 1971a. Rate of change of concomitantly variable codons. *J. Mol. Evol.* **1**, 84-96.

FITCH, W.M. 1971b. Towards defining the course of evolution: minimum change for a specific tree topology. *Syst. Biol.* **20**, 407-416.

FITCH, W.M. 1986. The estimate of total nucleotide substitutions from pairwise differences is biased. *Phil. Trans. R. Soc. Lond.* B **312**, 317-324.

FITCH, W.M. and F.J. AYALA. 1994. The superoxide dismutase molecular clock revisited. *Proc. Natl. Acad. Sci. USA* **91**, 6802-6807.

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. *Biochem. Genet.* **4**, 579-593.

FITCH, W.M. and J. YE. 1991. Weighted parsimony: does it work? In M.M. Miyamoto and J. Cracraft, eds. *Phylogenetic analysis of DNA sequences*, pp. 147-154. Oxford University Press, New York.

GILLESPIE, J.H. 1986. Variability of evolutionary rates of DNA. *Genetics* **113**, 1077-1091.

GILLESPIE, J.H. 1988. More on the overdispersed molecular clock. *Genetics* **118**, 385-386.

GOLDING, G.B. 1983. Estimates of DNA and protein sequence divergence: an examination of some assumptions. *Mol. Biol. Evol.* **1**, 125-142.

HAUSSLER, D., A. KROGH, I.S. MIAN, and K. SJÖLANDER. 1993. Protein modeling using hidden Markov models: analysis of globins. In T.N. Mudge, V. Milutinovic, and L. Hunter, eds. *Proceedings of the Twenty-Sixth Hawaii International Conference on System Sciences*. IEEE Computer Society Press, Los Alamitos, California.

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. *J. Classif.* **10**, 5-24.

HOLMQUIST, R. 1972. Empirical support for a stochastic model of evolution. *J. Mol. Evol.* **1**, 211-222.

HOLMQUIST, R., C. CANTOR, and T.H. JUKES. 1972. Improved procedures for comparing homologous sequences in molecules of proteins and nucleic acids. *J. Mol. Biol.* **64**, 145-161.

HOLMQUIST, R., M. GOODMAN, T. CONROY, and J. CZELUSNIAK. 1983. The spatial distribution of fixed mutations within genes coding for proteins. *J. Mol. Evol.* **19**, 437-448.

HUBBARD, T.J. and J. PARK. 1995. Fold recognition and ab initio structure predictions using hidden Markov models and β-strand pair potentials. *Proteins* **23**, 398-402.

HUGHEY, R. and A. KROGH. 1996. Hidden Markov models for sequence analysis: extension and analysis of the basic method. *CABIOS* **12**, 95-107.

JUKES, T.H. and C.H. CANTOR. 1969. Evolution of protein molecules. In H.N. Munro, ed. *Mammalian protein metabolism*, Vol. III, pp. 21-132. Academic Press, New York.

KARON, J.M. 1979. The covarion model for the evolution of proteins: parameter estimates and comparison with Holmquist, Cantor, and Jukes' stochastic model. *J. Mol. Evol.* **12**, 197-218.

KIMURA, M. 1980. A simple method for estimating evolutionary rates 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. 1983. *The neutral theory of molecular evolution.* Cambridge University Press, Cambridge.

KOONIN, E.V. and A.E. GORBALENYA. 1989. Evolution of RNA genomes: does the high mutation rate necessitate high rate of evolution of viral proteins? *J. Mol. Evol.* **28**, 524-527.

KROGH, A., M. BROWN, I.S. MIAN, K. SJÖLANDER, and D. HAUSSLER. 1994. Hidden Markov models in computational biology: applications to protein modeling. *J. Mol. Biol.* **235**, 1501-1531.

LANDER, E.S. and P. GREEN. 1987. Construction of multilocus genetic linkage maps in humans. *Proc. Natl. Acad. Sci. USA* **84**, 2363-2367.

LAWRENCE, C.E. and A.A. REILLY. 1990. An expectation maximization (EM) algorithm for the identification and characterization of common sites in unaligned biopolymer sequences. *Proteins* **7**, 41-51.

LOCKHART, P.J., A.W.D. LARKUM, M.A. STEEL, P.J. WADDELL, and D. PENNY. 1996. Evolution of chlorophyll and bacteriochlorophyll: the problem of invariant sites in sequence analysis. *Proc. Natl. Acad. Sci. USA* **93**, 1930-1934.

MARSHALL, C.R., E.C. RAFF, and R.A. RAFF. 1994. Dollo's law and the death and resurrection of genes. *Proc. Natl. Acad. Sci. USA* **91**, 12283-12287.

MCCLURE, M., C. SMITH, and P. ELTON. 1996. Parametrization studies for the SAM and HMMer methods of hidden Markov model generation. In *ISMB-96*, pp. 155-164. AAAI Press, St Louis.

MITCHISON, G. and R. DURBIN. 1995. Tree-based maximal likelihood substitution matrices and hidden Markov models. *J. Mol. Evol.* **41**, 1139-1151.

MIYAMOTO, M.M. and W.M. FITCH. 1995. Testing the covarion hypothesis of molecular evolution. *Mol. Biol. Evol.* **12**, 503-513.

NEI, M. 1991. Relative efficiencies of different tree-making methods for molecular data. In M.M. Miyamoto and J. Cracraft, eds. *Phylogenetic analysis of DNA sequences*, pp. 90-128. Oxford University Press, New York.

NEI, M. and T. GOJOBORI. 1986. Simple methods for estimating the numbers of synonymous and nonsynonymous nucleotide substitutions. *Mol. Biol. Evol.* **3**, 418-426.

PALUMBI, S.R. 1989. Rates of molecular evolution and the fraction of nucleotide positions free to vary. *J. Mol. Evol.* **29**, 180-187.

PENNY, D. 1974. Evolutionary clock: the rate of evolution of rattlesnake cytochrome *c*. *J. Mol. Evol.* **3**, 179-188.

PENNY, D., M.D. HENDY, and I.M. HENDERSON. 1987. 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. *Trends Ecol. Evol.* **7**, 73-79.

RABINER, L.R. 1989. A tutorial on hidden Markov models and selected applications in speech recognition. *Proc. IEEE* **77**, 257-285.

SCHÖNIGER, M. and A. VON HAESELER. 1994. A stochastic model for the evolution of autocorrelated DNA sequences. *Mol. Phylogenet. Evol.* **3**, 240-247.

STULTZ, C.M., J.V. WHITE, and T.F. SMITH. 1993. Structural analysis based on state-space modeling. *Protein Sci.* **2**, 305-314.

TAJIMA, F. and M. NEI. 1984. Estimation of evolutionary distance between nucleotide sequences. *Mol. Biol. Evol.* **1**, 269-285.

TUFFLEY, C. and M. STEEL. 1996. Modelling the covarion hypothesis of nucleotide substitution. Research Report No. 149, Department of Mathematics and Statistics, University of Canterbury. (*Mathematical Bioscience*, in press)

UZZELL, T. and K.W. CORBIN. 1971. Fitting discrete probability distributions to evolutionary events. *Science* **172**, 1089-1096.

WAKELEY, J. 1993. Substitution rate variation among sites in hypervariable region 1 of human mitochondrial DNA. *J. Mol. Evol.* **37**, 613-623.

WYCKOFF, H.W. 1968. Discussion. *Brookhaven Symp. Biol.* **21**, 252.