# Evolutionary analyses of large data sets: Trees and Beyond

A thesis presented in partial
fulfilment of the requirements

for the degree

of Doctor of Philosophy

in Mathematics at
Massey University

Barbara Ruth Holland
2001

# Abstract

The increasing amount of molecular data available for phylogenetic studies means that larger, often intra-species, data sets are being analysed. Treating such data sets with methods designed for small interspecies data may not be useful. This thesis comprises four projects within the field of phylogenetics that focus on cases where the application of current tree estimation methods is not sufficient to answer the biological questions of interest.

- A simulation study contrasts the accuracy of several tree estimation methods for a particular class of five-taxon, equal-rate, trees. This study highlights several difficulties with tree estimation, including the fact that some tree topologies produce "misleading" patterns that are incorrectly interpreted; that correction for multiple changes does not always increase accuracy, because of increased variance; and the difficulty of correctly placing outgroup taxa.

- A mitochondrial DNA data set, containing over 400 modern and ancient Adélie penguin samples, is used to estimate the rate of evolution. Straightforward tree-estimation is unhelpful because the amount of homoplasy in the data makes the construction of a single reliable tree impossible. Instead the data is represented by a network.

- A method, that extends statistical geometry, assesses whether or not a data set can be well-represented by a tree. The "tree-likeness" of each quartet in the data is evaluated and displayed visually, either for the entire data set or by taxon. This aids in identifying reticulate (or simply noisy) data sets, and also particular taxa that confound tree-like signal.

- Novel methods are developed that use pairwise dissimilarities between isolates in intra-species microbial data sets, to identify strains that are good representatives of their species or subspecies.

# Acknowledgements

First and foremost my thanks go to Michael Hendy and David Penny for sharing with me their wealth of ideas and passion for the subject of phylogenetics. If this thesis is at all readable, it is due to their tireless proof-reading efforts and insistence that I put forward my ideas in a semi-intelligible form.

I'd like to acknowledge the support of the many people who made this venture financially viable. Mike and David for the Marsden funded scholarship. The DAAD for making my first trip to Europe possible. Andreas Dress for supporting my stay in Bielefeld, Germany. Vincent Moulton and the STINT grant that made possible two trips to Sundsvall, Sweden. Allen Rodrigo for supporting a visit to Auckland.

I was most fortunate to have the opportunity to meet and work with a wide range of people over the last three years. Thanks to all my collaborators, you provided me with inspiration, interesting problems to work on and the benefit of your wide knowledge. In order of latitude, I express my gratitude to the Sundsvall crowd: Vincent Moulton and Katharina Huber for their fantastic hospitality, and Sverker Edvardsson for lending me his Athlon. In Greifswald I'd like to thank Dietmar Cieslik and Professor Kugelmann. In Bielefeld thanks to Andreas and Heidi Dress for their kind hospitality. Thanks also to Jan Weyer-Menkhoff for his heroic attempts to improve my Deutsch. Thanks to Jack Koolen, and in Duesseldorf, Bill Martin. Moving now to the southern hemisphere, in Auckland thanks go to Allen Rodrigo and Alexei Drummond for great advice on the "penguin chapter". Here at Massey it has been my great pleasure to work with David Lambert, Peter Ritchie and Jan Schmid, whose enthusiasm for penguins, penguins and microbes respectively, was highly infectious.

Thanks to two people who have been a great help on numerous matters, Peter Lockhart and Abby Harrison. Thanks also to the rest of the Thursday lunchtime gang, for sharing your work and for providing helpful comments on mine. I mustn't forget the maths grads, thanks for creating such a friendly atmosphere in which to

work.

A big thank you goes out to all of my friends, those who prevented me from going insane, *and* those that prevented me from going sane. To pick on a few by name, thanks to Agnieszka Szremska for persuading me that Maths and Biology made an interesting combination, and also for answering my dumb questions about genetics. Thanks to Maaike Bendall for always being a sympathetic ear and for your statistical know-how. Thanks to Paul Gardner for always being willing to bounce ideas, your handy computer hints and all those coffee breaks.

Behind everything I do there is always my family. Thanks Mum for your belief; Dad for the walks and talks at the beach; and my extended family for not being afraid to ask "so what is this phylogenetics stuff anyway?", and for providing food and shelter. Lastly thanks to my sister Miranda for proof reading, not complaining too vociferously when the dishes remained undone, and judicious application of hugs.

# Contents

# List of Figures

# List of Tables

# Chapter 1

# Introduction

## 1.1 Overview

Molecular phylogenetics is a young discipline. About fifty years ago, the structure of DNA was discovered by Watson and Crick, but it wasn't until the mid seventies that sequencing techniques were developed. Since then the amount of sequence data available has grown exponentially, this new data source enables biologists to ask a new range of questions about evolution and relationships between taxa. However, the sudden increase in the amount of data means that there has not been time to establish techniques for answering all these questions.

The problem in molecular phylogeny that has received the most attention, is that of resolving the phylogenies of data sets with small numbers of taxa, at the interspecies level; these were the types of data sets that first became available for study. Increased output of sequencing, and other types of molecular data, means that it is now feasible to have data sets with hundreds of taxa. Also, there are a growing number of intra-species data sets which it may not be appropriate to analyse with existing tree estimation methods.

This thesis describes four projects within the field of phylogenetics. These projects comprise chapters two through five. Broadly speaking chapter two is concerned with understanding existing tree estimation methods better; in chapters

three and five new techniques are developed to answer novel questions; chapter four describes a method of analysing molecular data sets, prior to tree estimation, to determine how successful, or appropriate a tree representation is likely to be.

The thesis is modular in nature, and it is intended that each of these chapters could be read stand alone. The aim of this introductory chapter is to bring out various linking themes that occur throughout, and also, to introduce some basic concepts and notation. First, a brief overview of the chapters. Readers unfamiliar with phylogenetic terminology might benefit from skipping ahead to section 1.3, which contains definitions of basic terms.

The topic of chapter two is a simulation study contrasting the accuracy of several tree estimation methods for a particular class of trees. The trees have five taxa, and equal rates of nucleotide substitution along all edges, that is, they are clock-like. Some interesting properties of tree estimation are noted, two are mentioned here: Firstly, we discover, that for some parameter combinations, trees which are consistently estimated can be chosen with far less frequency than each tree in a group of incorrect trees, for some bounded sequence lengths. This effect is observed in all methods, but occurs with the method parsimony [35] for a greater range of the parameter space. Secondly, it is found that correctly locating the outgroup is the hardest aspect of estimating these trees, sometimes it has a confounding effect, causing the ingroup taxa to form an incorrect tree, when trees without the outgroup would have been correctly recovered.

In chapter three, methods are developed to estimate the rate of evolution in Hyper Variable Region I (HVRI), a segment of the mitochondrial genome, in Adélie penguins. The data set comprises both modern, and dated ancient samples. This chapter includes a simulation study of the birth/death process in Adélie penguin populations, and the construction of a median network [3, 6]. A novel method is developed that uses median networks, of modern and ancient samples, to estimate the rate of evolution. This method is applied to the Adélie penguin median network.

Chapter four addresses the fundamental question of whether or not a data set

can be well-represented by a tree. An extension of statistical geometry is developed that assesses the "tree-likeness" of each quartet in the data and displays this information visually in a graph we call a $\delta$-plot. The potential of the $\delta$-plot method to identify reticulate data sets is explored. Furthermore, the method is extended to rank taxa in the data in order of how much they confound the tree-like signal, this may allow the identification of individual recombinant sequences.

Chapter five is somewhat removed from the phylogenetic framework of the other chapters. The aim here is to develop quantitative methods for selecting model strains that will be representative of their species or subspecies. The choice of model strains is of increasing importance in this genomic era, where individuals purported to be representative of their whole species, are completely sequenced, and analysed to determine, for example in the case of microbiological data, virulence factors and effective drug treatments.

Much of the work reported in this thesis is a result of collaborative projects. A diverse range of skills have been brought to bear on most of the problems described. One of the most enjoyable challenges in working in this area, is developing the ability to communicate well with people from many different scientific disciplines. These disciplines included mathematics, genetics, ecology, microbiology, statistics, and computer science. Notwithstanding the above comments, the work presented here is my own. In each chapter I indicate how my contribution fits within the overall scheme of each project, and also provide sufficient background material to set the scene.

Bound in at the back of the thesis are two papers on which I was a contributing author. The first paper [16] *Multiple Maxima for Maximum Likelihood in Phylogenetic Trees: An Analytic Approach* appeared in Molecular Biology and Evolution in 2000. It outlines an analytical technique for finding the maximum likelihood weights for some four-taxon trees. Many examples were found where the maximum likelihood surface for a given tree had multiple optima. This is of general importance, because most routines written to find the best maximum likelihood tree rely

on hill-climbing to optimise edge weights on the individual trees. My contribution to this project was to write numerical optimisation routines to confirm analytical results. Also, I identified many of the "interesting" data sets that lead to multiple optima.

The second paper [20] $\Delta$ *additive and* $\Delta$ *ultra-additive maps, Gromov's trees, and the Farris transform* was submitted to Discrete Applied Mathematics in July 2001. The paper extends known results for additive and ultra-additive metrics through the concept of $\Delta$ additivity. For a distance matrix $D$ on a taxa set $X$, $xy := d(x,y)$,

$$\Delta(D) := \max(uv + xy - \max(ux + vy, uy + vx)) : u, v, x, y \in X.$$

The paper also explores a tree construction method and corresponding bound of Gromov [40],

$$\|D - A\|_\infty \leq \Delta(D)\lceil \log_2(\#X - 1)\rceil,$$

where $A$ is the additive metric defined by the constructed tree. My contribution was to test the performance of Gromov's tree-construction and bound, on simulated data sets. I also helped to find worst-case example data sets on which Gromov's bound is tight.

## 1.2 Linking Themes

### 1.2.1 Trees and Beyond

I have attempted to suggest by my subtitle "Trees and Beyond", that the process of producing a single tree to reflect the historical relationships of some set of species, need not (and indeed should not) be the sole aim of phylogenetic analysis. A recurrent theme throughout this thesis is that of only using trees to the extent to which they are useful.

In the Adélie penguin mitochondrial data set of chapter three, trees provide a poor representation of the data, because they force arbitrary decisions about which mutations should be regarded as homoplasy. Networks, in contrast, allow one to display the many mutational pathways that might have occurred. Although, we do not expect that the Adélie mitochondrial sequences (which are maternally inherited [9]) have a genuinely reticulate history, the network models our uncertainty about which tree is correct.

Chapter four describes a preliminary analysis of distance data to determine how "tree-like" it is. One aim of this project is to identify those data sets for which a single tree can summarise the relevant information in the data, and those where it may be necessary to look "beyond" the tree for causes of other signal (or noise). For example, the method is used to identify reticulate data sets that may be better represented by networks, rather than trees.

In chapter five tree construction methods are used to produce a hierarchical clustering of the taxa set. Here, trees are useful insofar as they help to identify clustering within the data, but they only provide a first step towards answering the question of interest, "Which taxa are good choices as model strains?".

### 1.2.2 Large data sets

As the title indicates, the thesis is mainly be concerned with large data sets, where *large* refers to the number of taxa. Large is a relative term, in chapter two it may seem odd to claim that five taxa constitutes a large data set. However, comparable simulation studies that look in depth at the behaviour of methods on a single topology, such as [46], are with four-taxon trees. Also, we would expect that the five-taxon case studied here, will be embedded within larger problems. Most of the other data sets within the thesis would be considered large in a more genuine sense of the word.

The Adélie penguin data contains over 400 taxa in an alignment of 353 base-

pairs, of which the majority of sites are constant. Naturally, this puts a limit on the degree of resolution achievable in a tree. So, the strategy of using a network to represent this data is, in part, a response to the largeness of the data set.

Similarly, the *Candida albicans* data used as a case study in chapters four and five is large, with 266 taxa, as is *Helicobacter pylori* with 91 taxa. This made it computationally intractable to compare all possible sets of model strains, as there are $\binom{n}{k}$ ways of choosing $k$ model strains from $n$ taxa. To be useful for these data sets, the methods suggested for selecting model strains had to also be implemented as greedy approximations of the exact algorithms.

The methods in chapter four for measuring tree-likeness use statistical properties of the $\binom{n}{4}$ quartets in the data. Hence, these methods work better as $n$, the number of taxa, increases, because the effect of random noise on the properties being measured is reduced.

### 1.2.3 Using simulation as a tool

For all the projects described in this thesis, computers were an important tool. Computational power, combined with programming skills, made it possible to compare existing methods, and to test the effectiveness of newly developed methods.

Simulation studies are particularly useful when comparing the performance of different tree estimation methods under various models, as in chapter two. Given a weighted tree, a mechanism for nucleotide substitution, and the distribution of states at some arbitrary root, it is possible to simulate homologous sequences. Unlike using biological sequences, whose history is both unique and unobservable, testing methods on simulated sequences provides a repeatable experiment, where the output of a method can be compared to the known tree used to generate the methods input.

In chapter three the analytic results of Kingman's mathematical model for coalescence of populations [56, 55], are contrasted with simulation results under a

model that incorporates more realism. In this case, simulation allowed us to gather results about a complex situation that could not be analysed analytically.

In addition, computational power allows the use of a wide range of statistical tools including bootstrapping, jackknifing [27], randomisation tests and exact tests of significance [24, 69].

## 1.3 Basic Concepts

As far as possible I have tried to keep the notation consistent between chapters. Wherever encountered, $X$ is a set of *taxa*. A *taxon* can be a species, in the case of interspecies data, or an individual in an intra-species data set (taxa is the plural form of taxon). In some chapters the terms *isolate* or *strain* are used in place of the word taxon for intra-species data. $n$ is the number of taxa in a set $X$. Any subset of $X$ with four elements is called a *quartet*.

### 1.3.1 Trees and networks

A *graph* $G = (V, E)$, consists of a set of *nodes* $V$, and a set of *edges* $E$, $e \in E$ is of the form $e = \{x, y\}$ where $x$ and $y$ are in $V$. An edge $e = \{x, y\}$ is *incident* on the nodes $x$ and $y$. The *degree* of a node is the number of edges that are incident on it.

A *path* is an ordered set of nodes $(v_1, v_2, \ldots, v_k)$ where for each pair of nodes $(v_i, v_{i+1})$ there is an edge $e \in E$ with $e = \{v_i, v_{i+1}\}$. A *cycle* is a path that starts and ends at the same node. A graph is said to be *connected* if there is a path between each pair of nodes. A *tree* is a connected graph without cycles. A graph with cycles is called a *network*. Figure 1.1 (a) is an example of a connected graph with cycles.

A node of degree one is called an *external* node or *tip*. A node of degree greater than one is an *internal* node. In figure 1.1 (b), node $z$ is an internal node with degree 3, whereas node $x$ is an external node with degree 1. In a *phylogenetic tree* the external nodes are labelled with the taxa set $X$. Note that an external

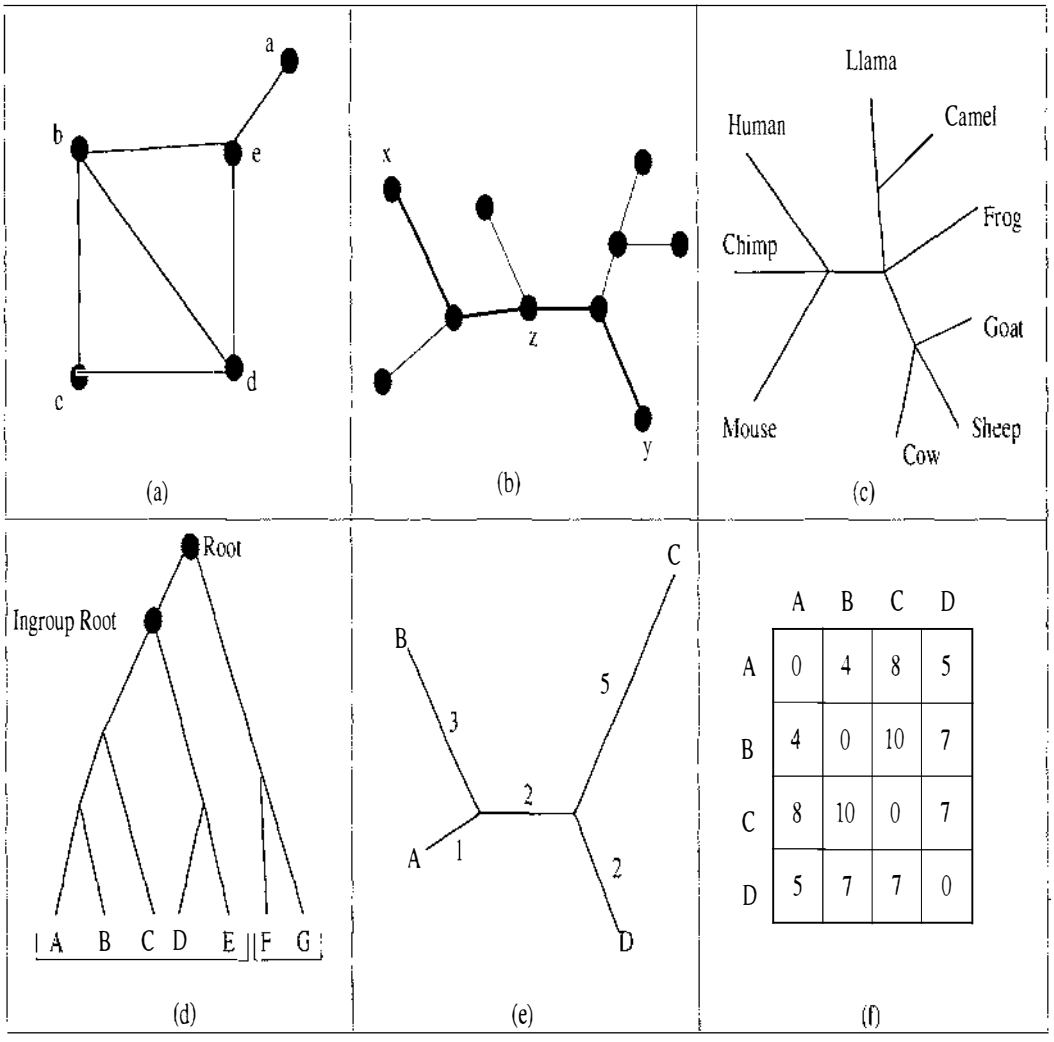


Figure 1.1: Examples of basic concepts with graphs and trees. Panel (a) shows a *graph* with 5 *nodes* and 6 *edges*. The ordered set of nodes $(b, e, d, c, b)$ is a *cycle*. Panel (b) shows a *binary* tree, the unique *path* between nodes $x$ and $y$ is shown by the bold line. Node $x$ is an *external node* as it has degree 1, whereas $z$ is an *internal node* with degree 3. Panel (c) shows a *tip-labelled*, or *phylogenetic, multifurcating* tree. Panel (d) shows a *rooted, binary* tree. The taxa set {F, G} is an *outgroup* to the *ingroup* taxa set {A, B, C, D, E}. The taxa A, B and C are an example of a *clade*, but A, B and D are not. Panel (e) shows a *weighted tree*. The metric induced by this tree on the taxa set A,B,C and D is shown in panel (f).

node may be labelled by more than one taxon. Throughout, I will often refer to phylogenetic trees simply as trees, or sometimes as tip-labelled trees. Panels (c), (d) and (e) of figure 1.1 all display phylogenetic trees.

Trees in which all nodes have degree one or three are called *binary* trees, or *fully resolved* trees. Trees where some nodes have degree greater than three are called *multifurcating* or *partially resolved.* In figure 1.1 (b), (d) and (e) are binary trees, and (c) is a multifurcating tree.

In a *weighted* tree each edge has a specified weight (or length), in phylogenetic trees weights are non-negative, real numbers. A weighted phylogenetic tree induces a pseudo-metric on the taxa set $X$ which labels its external nodes. (Throughout the thesis I simply refer to metrics rather than pseudo-metrics, although $d(x, y) = 0$ with $x \neq y$ is possible.) Each pair of taxa $x, y \in X$ is connected by a unique path in the tree. The sum of the edge lengths (weights) along the path gives the distance between the pair of taxa. A metric that can be represented by a weighted tree in this way is called *additive.* The additive metric in figure 1.1 (f) corresponds to the tree in panel (e).

A *directed* edge is an ordered pair of nodes. A *rooted* tree has a node designated the root, and all edges in the tree are directed away from the root. In a phylogenetic tree the root represents a common ancestor of the taxa set $X$, and each internal node represents a common ancestor to the nodes it separates from the root. Figure 1.1 (d) is a rooted tree. In a *rooted binary tree* the root node has degree two, and all other nodes have degree one or three.

A weighted, rooted tree, where the distance (induced by the edge weights) is the same from each taxon to the root, is called *ultrametric*, or *clock-like* when the weights on the edges correspond to units of time.

In an unrooted phylogenetic tree, a root can be identified by the inclusion of an additional taxon, or group of taxa called an *outgroup.* The taxa that do not belong to the outgroup are called the *ingroup.* The internal node where the outgroup joins the unrooted tree is designated the root of the ingroup tree, this is indicated, in

figure 1.1 (d), by the node labelled "ingroup root".

A *split* of the taxa set $X$, divides $X$ into two non-empty disjoint subsets, say $A$ and $B$, whose union $A \cup B = X$. A split is written $A|B$ $(= B|A)$. Each edge $e = \{x, y\}$ in a tree induces the split $A|B$ of the taxa set $X$, where removing $e$ leaves two connected subtrees on taxa sets $A$ and $B$.

Two splits, $A|B$ and $C|D$, of a taxa set $X$, are said to be *compatible* if at least one of the sets $A \cap C$, $A \cap D$, $B \cap C$, or $B \cap D$ is the empty set. This is equivalent to there existing a tree with edges $e_1$ and $e_2$, where $e_1$ induces $A|B$, and $e_2$ induces $C|D$. Splits that are not compatible are *incompatible*.

In a rooted tree, the set $A$ or $B$ in the split $A|B$, that does not label the subtree that includes the root, is called a *clade*.

### 1.3.2 Tree estimation methods

A *tree estimation method* is an algorithm that takes as input a data set, and returns a phylogenetic tree. Where there is no confusion, I refer to these simply as methods. A *dissimilarity based method* is a method where the input data is in the form of a *dissimilarity matrix*. A *sequence based method* is a method where the input data is in the form of a *sequence alignment*.

Methods typically consist of three components,

- an optimality criterion,
- a search policy, and
- a correction policy.

An *optimality criterion* assigns a score to a tree given the data. For example, with the method parsimony, the score for a tree is the minimum number of substitutions required to account for the data on that tree. With the method of maximum likelihood, the score for a tree is the probability of observing the data, given that tree and some model of nucleotide substitution.

The *search policy* dictates how the method searches through the space of all trees. Strategies include: exhaustive search, either by brute force or branch and bound, and heuristic search where an initial tree is constructed and improved upon until a locally optimal tree is found. Exhaustive search quickly becomes intractable as the number of taxa, $n$, gets large, because the number of unweighted trees on $n$ taxa is $(2n-5)!! = \Pi_{i=1}^{n-2}(2i-1)$, for $n \geq 3$. For example, the number of trees on $n = 8$ taxa is $11 \times 9 \times 7 \times 5 \times 3 \times 1 = 10395$.

Methods such as neighbor-joining and UPGMA, which are based on dissimilarity data, combine the optimality criterion and the search policy into a greedy constructive heuristic. At each stage a criterion determines which taxonomic units should be amalgamated into a single unit. The dissimilarity matrix is then updated, and the process repeats until all units have been joined.

Given two aligned sequences the number of observed differences in the sequences will be a lower bound on the actual number of substitutions that occurred in the evolutionary path connecting them. This is due to the possibility of multiple substitutions at the same site. For instance, a nucleotide A might mutate to a C and then to a G, but this would be seen as only one change; or alternatively, an A may change to a G and then back to an A, which would be seen as no change at all. For a chosen model of nucleotide substitution it is possible to estimate the actual number of substitutions that occurred given the number of observed differences. The *correction policy* of a method dictates if a correction will be used, and if so according to which model.

The main desirable property of a tree estimation method is that it be *accurate*, in other words, that it outputs the correct tree. The correct tree in this context means the tree which displays the evolutionary relationships amongst the taxa (if indeed it is a tree rather than a network). In general it is considered more important that the correct unweighted tree is recovered, than that the edge weights are proportional to the times of speciation events.

Except in the occasional controlled laboratory experiment, the correct tree re-

flects an unobservable, unrepeatable historical event. For this reason the accuracy of methods is typically measured through simulation studies. These have the twin virtues that (i) the correct tree, i.e. the one used to generate the data, is known, and (ii) experiments are repeatable. Accuracy is defined as the probability that the method will return the correct tree, and is estimated by the proportion of simulation runs in which the correct tree is recovered. An alternative definition is the average number of edges in the tree correctly recovered. In the simulation study in chapter two the first definition is used. If a method has higher accuracy than another method for fixed sequence lengths, then it is said to be the more *powerful* method.

Obviously, no method will be correct all the time due to sampling error introduced, for example, by short sequence lengths. However, it is a desired property of a method, that as the sampling error tends to zero, the probability of the method outputting the correct tree tends to one. This property is known as *consistency.*

### Input Data

The two forms of input used by the methods dealt with in this thesis are *sequence alignments* and *dissimilarity matrices.*

A sequence alignment is an $n \times c$ matrix of character states over a finite alphabet, for example {A,C,G,T}, where each row corresponds to the sequence of a different taxa, and the columns are *homologous* sites (sometimes called characters), that is, they are assumed to have evolved from a common ancestral character state. Sequence length is denoted $c$, so a sequence alignment will have $c$ columns. A site that has the same character state for all taxa is called *constant.* A site in which all taxa but one share have the same character state is called a *singleton.* An example sequence alignment in given in table 1.1. In biological data the character states often come from the alphabet A,C,G,T,-,?, where '-' indicates an insertion or deletion (*indel*), and '?' indicates that the character state is not known. All the simulated sequences used in the thesis are over the four letter alphabet {A,C,G,T}.

| Site | 1 | 2 | 3 | 4 | 5 |
|---|---|---|---|---|---|
| Kiwi | A | A | C | C | C |
| Weka | A | A | C | G | G |
| Huia | A | G | C | C | A |
| Tui | A | G | A | C | A |

Table 1.1: An example sequence alignment of length $c = 5$ for $n = 4$ species. Site 1 is constant, sites 3 and 4 are singleton.

---

The dissimilarity matrices are symmetric, square, $n \times n$ matrices, with zero diagonal and non-negative real-valued entries.

DEFINITION: 1 (DISSIMILARITY) *A dissimilarity* $D : X \times X \to \mathbb{R}$ *satisfies the conditions:*

1. $d(x, y) \geq 0$, $d(x, x) = 0$, *for all* $x, y$

2. $d(x, y) = d(y, x)$, *for all* $x, y$

There are cases where biological data does not satisfy these properties, for instance, in DNA-DNA hybridisation data one taxa may not hybridise perfectly with itself, implying $d(x, x) > 0$. All the input data used within this thesis satisfies this definition of a dissimilarity.

If a third property, (the triangle inequality)

$$d(x, y) + d(y, z) \geq d(x, z) \text{for all } x, y, z \in X$$

holds, then $D$ is a *distance metric.* Note, that neither dissimilarities or distance metrics require that $d(x, y) > 0$ for $x \neq y$; in biological data it will sometimes be the case that different taxa are identical on the characters measured.

The dissimilarity matrices in this thesis are obtained from two sources. Firstly, they are derived from sequence alignments by taking *Hamming distances.* The Hamming distance between a pair of aligned sequences is defined as the number of differences in their sequences divided by the length of the sequences $c$. Table 1.2

| | Kiwi | Weka | Huia | Tui |
|---|---|---|---|---|
| Kiwi | 0 | 0.4 | 0.4 | 0.6 |
| Weka | 0.4 | 0 | 0.6 | 0.8 |
| Huia | 0.4 | 0.6 | 0 | 0.1 |
| Tui | 0.6 | 0.8 | 0.1 | 0 |

Table 1.2: An example of a distance matrix. It contains the Hamming distances for the sequence alignment shown in table 1.1.

gives an example distance matrix containing the Hamming distances between the sequences displayed in table 1.1. Hamming distances always satisfy the triangle inequality.

The second source of dissimilarity data is Restriction Fragment Length Polymorphism (RFLP). RFLP, works by cutting DNA with restriction enzymes and then separating the resulting DNA fragments by running them through an electrified field on a gel. Different sized fragments travel at different speeds down the gel producing a pattern of bands. The restriction enzymes are specific to certain short substrings of DNA, so if one sequence has insertions, deletions, or point mutations relative to another it may be cut in different places, producing its own distinctive spectrum of fragment sizes. The gels are aligned and scored according to the strength of bands of a particular molecular weight.

A similar source of dissimilarity data called Amplified restriction Fragment Length Polymorphism (AFLP) is used in chapter 5. Fragments of DNA that match particular primers are amplified by the polymerase chain reaction. These amplified fragments are then separated by running them through an electrified field on a gel.

The scheme used by my microbiologist collaborators to compute dissimilarities from this type of data is

$$d(a,b) = 1 - \frac{\sum_{i=1}^{B}(a_i + b_i - |a_i - b_i|)}{\sum_{i=1}^{B}(a_i + b_i)}, \quad a, b \in X,$$

where $B$ is the number of bands, and $x_i \in \{0, 1, 2, 3\}$ is the intensity of band $i$ in

isolate $x$ [86]. Alternatively the bands can be coded for presence or absence forming a binary character matrix, and Hamming distances taken between the rows.

### 1.3.3 Models of character substitution

A model of character substitution specifies the rates at which each character mutates (changes) into another character. Models are used in a variety of contexts:

- They are used as part of the method maximum likelihood to evaluate the probability of a weighted tree generating the observed data.
- They are necessary in order to generate simulated data, where data is generated on a weighted tree, from a known sequence at the root, according to some specified model of character substitution.
- They are used by any method that incorporates a correction for multiple substitutions to estimate the actual number of substitutions that occurred between two taxa given the observed number of substitutions.

The two-state symmetric model, one of the models used in chapter two, has the following rate matrix:

$$Q = \begin{bmatrix} -x & x \\ x & -x \end{bmatrix}, \quad 0 \leq x$$

The transition matrix for an edge of length $t$ is $P = e^{Qt}$, so $P$ is of the form:

$$P = \begin{bmatrix} 1-y & y \\ y & 1-y \end{bmatrix}, \quad 0 \leq y \leq \frac{1}{2}$$

The $(i, j)$ entry of $P$ gives the probability of a substitution from state $i$ to state $j$ after a time $t$. All entries in the transition matrix are in the range $[0, 1]$, and the rows and columns must sum to one. This forms a Markov process, the future character state at a site depends only on the current state, not on the site's history.

# Chapter 2

# Tree estimation with equal rates: A simulation study with five taxa

## 2.1 Introduction

All tree estimation methods can be inconsistent if incorrect assumptions are made about the mechanism of evolution [46, 96]. Conversely, with the appropriate correction for multiple substitutions at sites, most methods are consistent. However, corrections for multiple substitutions increase the variance of distance estimates [57, 96], so it does not follow that the appropriate correction will always lead to more accurate estimates of the tree for bounded sequence lengths. Simulation studies provide a framework in which to study the interaction of these effects.

The best known example of inconsistency is for maximum parsimony on the four-taxon case with unequal rates, this has been extensively studied [32, 46, 100]. However, it was shown by Hendy et al. [43] that unequal rates are not necessary for parsimony to be inconsistent. In this chapter I focus exclusively on the five-taxon topology, with equal rates, for which parsimony can be inconsistent.

A simulation study is conducted under both the symmetric two-state model [31] and Kimura two-parameter (four-state) model [54]. I compare the accuracy of several popular phylogenetic methods on this topology, including parsimony, neighbor-

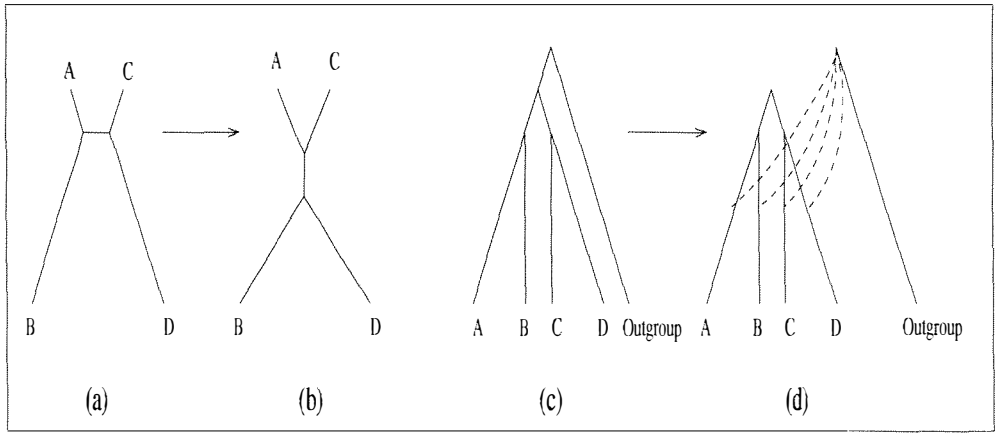


Figure 2.1: Common errors in tree construction. In the Felsenstein zone (a), many methods recover tree (b). The tree (c) is the equal rates topology on five taxa for which uncorrected parsimony can be inconsistent, (d) shows the most common type of error made in reconstructing trees from data generated on (c), the outgroup joins with equal probability to taxa A, B, C or D.

joining, UPGMA, maximum likelihood, and split decomposition. The effectiveness of performing corrections for multiple substitutions given various sequence lengths is explored for neighbor-joining and parsimony.

A new problem of tree estimation methods is observed and is described as a "misleading zone". Within this zone a method is consistent, but despite this, for bounded sequence lengths some incorrect trees are estimated more frequently than the correct tree.

In addition, we use this five-taxon example with equal rates to study the effectiveness of using an outgroup to root an ingroup tree. Rooting the tree is known to be one of the more difficult aspects of tree estimation [62, 66, 88]. It can even have a confounding effect, causing the ingroup taxa to form an incorrect tree when trees without the outgroup would have been correctly recovered.

## 2.2 Background

In general the problems of accuracy and consistency have been well studied for four-taxon trees. This was motivated by the discovery of the "Felsenstein zone" in which maximum parsimony was found to be inconsistent [32]. For a tree in the "Felsenstein zone", sequences do not evolve at equal rates along the edges; the most frequently studied case has a long and a short edge on each side of the short internal edge (figure 2.1 (a)). The typical error made in reconstructing a tree from sequences evolved under this model is to join the long edges together (figure 2.1 (b)). This effect has been studied both analytically [32, 43], and by simulation [46, 100]. Penny et al. in [74] (pg 173) report that for two-state data, simple clustering procedures that, at each stage, join the two taxa with smallest uncorrected distance, will fail to converge under the same conditions that uncorrected parsimony fails. Hillis et al. [46] found in their simulation study that distance based methods were inconsistent for this type of tree, unless the distances had been corrected for multiple changes.

In this chapter we study another class of trees that has received less attention, trees on five taxa that evolved with equal rates, that is, under a molecular clock. While equal rates are sufficient to guarantee consistency of neighbor-joining and UPGMA on observed distances, this is not true for parsimony; it can be inconsistent on these trees when the internal edges are sufficiently short compared to the external edges. With five taxa this only occurs on the rooted tree where the fifth taxon is an outgroup linked to the internal edge of the four taxon subtree [43] (figure 2.1 (c)). The most common type of error made in estimating trees from data evolved along this model is for the outgroup to be paired with one of the four external edges (figure 2.1 (d)). Trees with this type of structure are observed in real biological data, see for example, Lockhart et al. [60].

Even when the internal four taxon tree is well away from the "Felsenstein zone" the inclusion of an outgroup can cause the method of maximum parsimony to become inconsistent. This problem was discovered by Hendy and Penny [43], and

is further discussed in Steel et al. [93]. One surprising observation in this case, is that maximum parsimony is consistent on each subset of four taxa (each quartet), but not on the complete five-taxon tree.

Previous simulation studies have found that the accuracy of tree estimation methods is better when there are equal rates of change along the edges of the tree [83, 91, 92]. By using a model five-taxon tree that obeys the molecular clock, we remove this potential source of inaccuracy and focus on the effect of having a combination of short and long edges.

With consistent methods there is still the question of what length of sequence is required to give a "reasonable" chance of recovering the correct tree. We give examples of trees on which parsimony is consistent, but for simulated sequences of fixed length, specific incorrect trees are each recovered with significantly greater frequency than the correct tree.

We also use the five-taxon example to investigate the effectiveness of using an outgroup to root a tree. The inclusion of an outgroup can affect the accuracy of tree estimation. Phylogenetic methods can be misled by sampling error or homoplasy when the root of the ingroup is on a short internal edge, as shown in figure 2.1 (c) and (d). This effect has been termed "long edge attraction" [43], it can lead to estimating trees with correct ingroup structure but an incorrect order of divergences. Incorrectly assigning the root will change which groups of taxa are deemed to be monophyletic, and which character states are identified as ancestral.

There are also cases where the ingroup by itself is recovered correctly (figure 2.2 (a)), but inclusion of an outgroup causes the method to change its estimate of the ingroup tree resulting in it becoming incorrect (figure 2.2 (b)).

To illustrate the importance of the problem of identifying the position of the root, note that locating the root is currently a major problem for birds, for mammals, and for flowering plants. The origin of the birds is contentious, commonly the root is placed between paleognaths (ratites and tinamous) and neognaths (all other birds). In contrast, the study by Mindell et al. [67] of mitochondrial genomes places

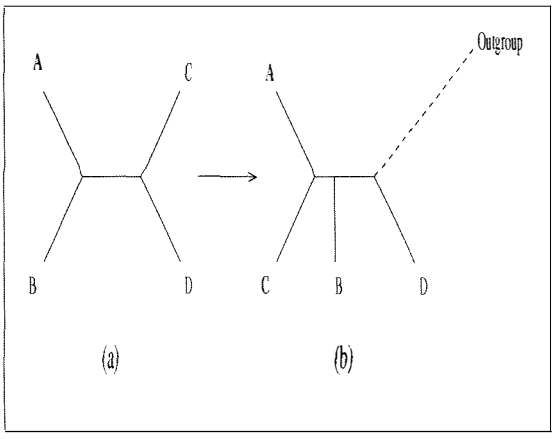


Figure 2.2: Sometimes the inclusion of an outgroup can cause the correct ingroup structure (a) to become incorrect (b). This is an example of the outgroup **confounding** tree estimation.

the root within the song-birds. However, this could be a long edge attraction problem, as the phylogeny reported by Cooper et al. [17], suggests that song-birds are on an edge with a faster than average rate of change. Soltis et al. and Qui et al. [90, 77] have both studied the origins of the flowering plants and found that *Amborella*, a shrub found only in New Caledonia, is sister taxa to all other angiosperms. This is another case where outgroup placement may have been "attracted" to a long edge. Waddell et al. [103] use mitochondrial DNA to study the phylogeny of eutherian mammals; although they found the hedgehog is a possible sister taxa to the other eutherians, they point out that the result is suspect for the reasons described above. Despite its biological importance the problem of accurately identifying the root of a tree has not been extensively studied.

## 2.3 Models and Methods

We generated sequences under a molecular-clock model of nucleotide substitution on the five-taxon tree (illustrated in figure 2.1 (c)). Two different models of nucleotide substitution were used, the symmetric two-state model [31], and the four-state

Kimura two parameter model (K2P) [54]. A range of tree estimation methods were applied to each sample sequence; the tree each produced was compared to the generating tree. The models of sequence evolution, the generating tree, and the tree estimation methods are each outlined in detail below.

### 2.3.1 Models of sequence evolution

Initial simulations used the symmetric two-state model [31]. This model assumes equal base frequencies at the root and a single rate of substitution across all sites, with all sites independent and free to vary.

The Hadamard conjugation (equation 2.1) [45] was used to calculate the expected sequence patterns, $\mathbf{s}$, as functions of the edge lengths $\mathbf{q}$, on the tree.

$$\mathbf{s} = H^{-1} \exp(H\mathbf{q}), \tag{2.1}$$

where $H = H_4$ is the $16 \times 16$ symmetric Hadamard matrix with $\pm 1$ entries, constructed recursively from $H_0 = [1]$, by $H_{n+1} = [H_n, H_n; H_n, -H_n]$, for $n \geq 0$. For a full description of Hadamard conjugation in this context see Hendy et al. [45].

In general, for studies with $n$ taxa, the computational cost of generating $\mathbf{s}$ through the Hadamard transformation is $\bullet(n2^n)$ [15]. However, this only needs to be done once for each weighted tree. Sampling from $\mathbf{s}$, which in simulation studies such as this is repeated many times for each weighted tree, is computationly cheap ($O(nc)$, where $n$ is the number of taxa and $c$ is the sequence length). For small taxa numbers, such as $n = 5$, it is faster to generate two-state character sequences using the Hadamard method than by simulating them along a tree. For a more detailed discussion of the relative computational costs see Charleston et al. [15].

Reversals and parallel changes are more likely to occur with two-state characters than with four. Indeed, with two states any even number of changes at a site will be seen as no change at all. For comparison and increased biological realism, further simulations were conducted using the four-state K2P model [54]. The K2P model

assumes equal base frequencies at the root, and specifies both a transition rate and a transversion rate. For these simulations the transition/transversion rate ratio was set at $\kappa = 2$. All sites are independent, identical, and free to vary.

### 2.3.2 Tree estimation methods

| Notation | Method |
|---|---|
| UP | UPGMA applied to observed distances. |
| NJ | Neighbor-joining applied to observed distances. |
| CNJ | Neighbor-joining applied to distances corrected for multiple changes. (Unless otherwise stated the standard correction was applied.) |
| MP | Maximum parsimony applied to the observed sequence spectrum ($\hat{\mathbf{s}}$). |
| CMP | Corrected parsimony (applied to the conjugate spectrum, $\hat{\mathbf{q}} = H^{-1}\ln(H\hat{\mathbf{s}})$). |
| ML | Maximum likelihood using the appropriate two or four-state model. |
| ST | Split decomposition (a.k.a. *SplitsTree*). |

Table 2.1: The following notation for tree estimation methods is used throughout the chapter. For a detailed description of these methods consult Swofford et al., (1996) [96]. $H$ is the Hadamard matrix $H_4$ [45], described above.

The tree estimation methods covered in this study are shown in table 2.1. A tree-estimation method is considered to consist of three separate components [73, 93]:

- the optimality criterion,
- the policy used to correct for multiple substitutions, and
- the search strategy.

For example, in this simulation CMP is the parsimony optimality criterion applied to data that has been corrected for multiple changes via Hadamard conjugation[1] [93], with an exhaustive search strategy. With NJ, two different corrections were applied: the standard distance correction for the two-state model (described in

[1] Correcting parsimony for multiple changes is readily done by the Hadamard conjugation. The correction can be based on any of the simpler models of evolution such as the two-state model, the Jukes Cantor model and the Kimura 2P and 3ST models. It can also include corrections for models with a gamma distribution of rates across sites [105].

Swofford et al. [96]); and the reduced bias correction of Tajima [98]. As the number of unweighted trees on five taxa is only fifteen, MP, CMP and ML used exhaustive search strategies. NJ, CNJ and UP are stepwise clustering algorithms, their search strategies are greedy (there is no opportunity to backtrack after a decision to amalgamate two taxonomic units has been made).

I developed code for most of the algorithms myself, which, where possible, was verified against existing software (*phylip* [34], *PAUP** [97] and *SplitsTree* [49]). With the exception of ML the algorithms are not complicated to encode, the advantages of designing one's own code include:

- Gaining a better understanding of the algorithms.

- Streamlining the algorithms for the five-taxon case. For example, there is no need to use a branch and bound procedure with MP for five taxa, or, with two-state data, even to use the Fitch algorithm. One simply checks which of the 15 possible trees on five taxa has the largest number of sites compatible with its internal edges.

- Incorporating a random tie-breaking procedure at each decision making step. For example, if UP has three equally good options for amalgamating taxonomic units at a given step, then each possible choice is assigned a probability 1/3 of being selected. This is very important for providing unbiased long-run statistics on the accuracy of the methods. Shuffling the input order of taxa before applying an algorithm is not sufficient to remove bias in estimating trees. If UP is implemented by adding newly amalgamated taxonomic units to the bottom of the distance matrix then, without tie-breaking, there will be a bias for joining either single taxa or grouped taxa. The well-known phylogeny package *phylip* [34] suffers from this problem, it only gives the option of shuffling the taxon order rather than using random tie-breaking at each step.

| Edge($\alpha$) | Expected number of substitutions per site ($q_\alpha$) |
|---|---|
| $1\|2,3,4,5$ | 0.1 |
| $2\|1,3,4,5$ | 0.1 |
| $3\|1,2,4,5$ | 0.1 |
| $4\|1,2,3,5$ | 0.1 |
| $1,2\|3,4,5$ | $x = 0.01$ to 0.1 in steps of 0.01 |
| $3,4\|1,2,5$ | |
| $1,2,3,4\|5$ | $y = 0.2$ to 0.4 in steps of 0.02 |

Table 2.2: The weights on the generating tree $T$ (see figure 2.3). The vector of the above values, indexed by all the possible splits $\alpha$ induced by an edge of $T$, is called the *edge length spectrum* $\mathbf{q}$. The notation $A|B$ denotes the edge whose removal from the tree partitions the leaf set into two subsets, one containing the set of taxa A and the other the set B. For example $1|2,3,4,5$ indicates the external edge ending in taxon 1.

- Designing the input and output format to be convenient for the particular requirements of the simulation. For example, the output of the *SplitsTree* program is a graphical display of a network. All that we require to check the accuracy of the method are the isolation indices for the ten non-trivial splits.

I wrote my own code for ML on two-state sequences, but for four-state sequences, I used the *phylip* package *dnaml* [34].

### 2.3.3 The tree and generated sequence alignments

The model tree that was used in all of the simulations is shown in figure 2.3 and described in terms of the splits ($q_\alpha$) it generates in table 2.2. For each weighted tree described by this range of values, homologous sequences were generated.

In the two-state model simulations we use the Hadamard conjugation [45] to calculate the corresponding expected sequence spectrum $\mathbf{s}$,

$$\mathbf{s} = H^{-1} \exp(H\mathbf{q}).$$

The sequence spectrum $\mathbf{s}$ is indexed by the splits $\alpha$. The component $s_\alpha$ is the probability of observing the split $\alpha$ at a particular site. Monte Carlo sampling

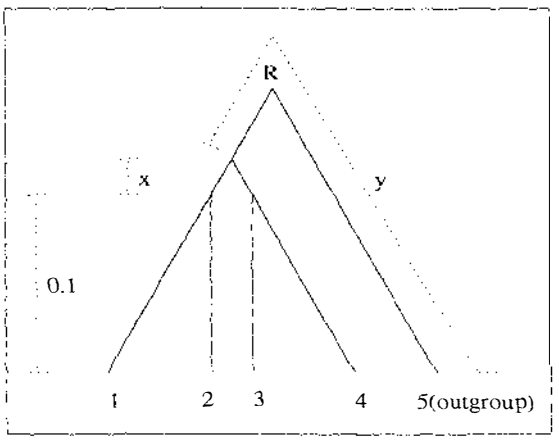


Figure 2.3: The generating tree (((1,2),(3,4)),5). The value of x (internal edge length) ranges from 0.01 to 0.10 (in steps of 0.01) and the value of y (outgroup edge) ranges from 0.2 to 0.40 (in steps of 0.02. The units are the expected number of substitutions per site. Taxon 5 is an outgroup to the other taxa. The root R is fixed at the mid-point of the path between taxa 1 and 5, the path length from the root R to each tip is $(y + x + 0.1)/2$, so the molecular clock is satisfied.

from the sequence spectrum $\mathbf{s}$ provides the sample sequence data, $\hat{\mathbf{s}}$. A sample of size $c$ is equivalent to a sequence of length $c$. For the four-state model it is possible, although more complicated, to use Hadamard conjugation for sequence generation. Hadamard conjugation is more computationally expensive for four states as the sequence spectrum is of the length $(2^{n-1})^2$ rather than $2^{n-1}$ for two states, e.g. 256 instead of 16, for $n = 5$. Instead I chose to use the program *Seq-gen* [81] to obtain sequence samples. The sample sizes (sequence lengths) used were $c = (100, 200, 500, 1000)$.

### 2.3.4 From sequences to distances

The sequences were converted into Hamming distances for UP, NJ, CNJ and ST. There is the option of correcting these distances for multiple changes, although, not all sets of data are correctable by the standard method (as given in Swofford et al. (1996) [96]). The correction assumes equal rates of substitution between all

pairs of bases, the formula for $r$-state data is

$$\hat{d}_{ab} = -\frac{r-1}{r}\ln(1 - \frac{r}{r-1}\frac{k_{ab}}{c}),$$

where

$k_{ab}$ = number of differences between sequence $a$ and sequence $b$,

$c$ = sequence length,

$r$ = the number of character states.

If the uncorrected distance $\frac{k_{ab}}{c}$ is greater than $\frac{r-1}{r}$, the argument is negative and the log is undefined over the real numbers. Although the expected value of $\frac{k_{ab}}{c}$ is less than $\frac{r-1}{r}$, some observations are possible when the variance is high, as occurs with small values of $c$. Data sets where this occurred were not used by the CNJ method. In the two-state simulations this resulted in rejecting about 5% of the samples for $c = 100$ and about 1% for $c = 200$. For $c \geq 500$ the distances observed were always less than $\frac{r-1}{r} = 0.5$ $(r = 2)$.

The reduced bias correction of Tajima [98] is always applicable. The correction for two-state characters is given by

$$\hat{d}_{ab} = \sum_{i=1}^{k_{ab}} \frac{k_{ab}^{(i)} \times 2^{i-1}}{i \times c^{(i)}},$$

where

$$z^{(i)} = \frac{z!}{(z-i)!}.$$

I wrote my own code to implement this correction in the two-state case.

For the four-state simulations I used the *phylip* package *dnadist* [34] to do the K2P distance correction, see Swofford et al. (1996) [96], pg 456, or the *phylip* documentation, for details. The Tajima reduced bias correction is not supported by *dnadist*.

### 2.3.5 The simulation process

All of the sample data was generated in advance, so each method had the same input (in the appropriate distance or sequence format). The trees estimated by each method were compared to the model tree. Accuracy is defined here as the frequency with which the unweighted generating tree was recovered. In the two-state character simulations, for each value of the parameters, the simulation process was repeated 10,000 times for every method other than ML, where it was repeated 1000 times. In the four-state character simulations the process was repeated 1000 times for every method other than ML, where it was repeated 100 times. The simulation process is summarised in figure 2.4.

The results of the simulation are discussed in section 2.4.1 and summarised in table 2.3. Section 2.4.2 describes and explains an effect we call a "misleading zone", in which although a method is consistent, for bounded sequence lengths, a group of incorrect trees are each selected more frequently than the correct tree. Section 2.4.3 contains a detailed analysis of the classes of error that can occur.

## 2.4 Results

### 2.4.1 Accuracy of the Methods

In general, the results found using the two-state and four-state models show similar trends. I firstly discuss the points of interest that are common to the results for both models, then highlight the differences that were observed.

#### General effects

In figure 2.5 the results for the NJ method with $c = 100$ and four-state sequences are given to illustrate how the results in this section are presented. Throughout the section, accuracy of tree estimation is indicated by a colour scale that ranges from deep blue (low accuracy) to red (high accuracy).

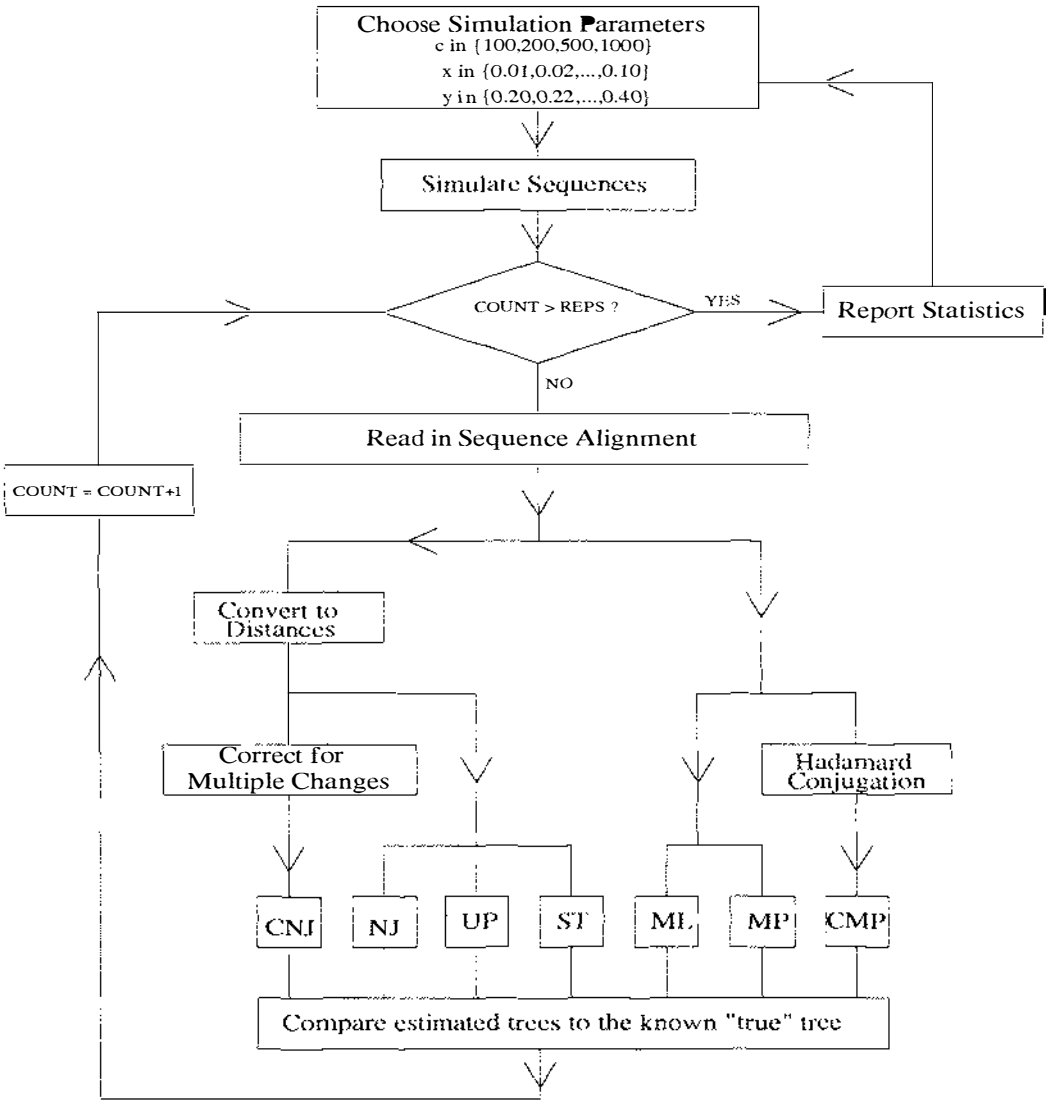


Figure 2.4: Flowchart of the simulation process. For the two-state simulations the number of repetitions was 10,000, for four states it was 1,000. Note that fewer repetitions were performed for ML (1000 and 100 respectively).

Figure 2.6 contains the results for two-state sequences, and 2.7 gives the results for four-state sequences. As expected, all methods are less accurate when the internal edges are short and the outgroup edge is long. For consistent methods, when the sequence length $c$ becomes long enough, this problem is overcome. However, MP is inconsistent for some combinations of parameters where $x$ is small and $y$ is large. For these trees that are inconsistently estimated by MP, as sequence length

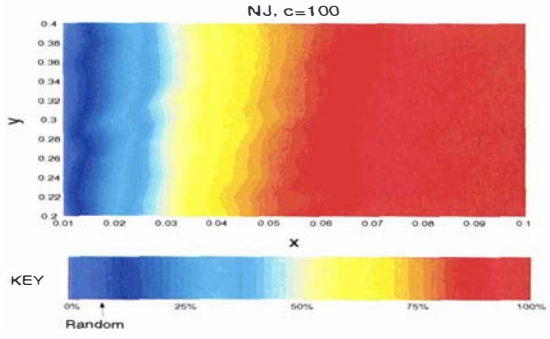


Figure 2.5: An example, using the results for NJ with $c = 100$ and four-state sequences, to demonstrate how the results in figures 2.6, 2.7 and 2.8 are presented. The horizontal axis shows the length of the internal edges (0.01–0.10), and the vertical axis is the length of the outgroup edge (0.20–0.40) . The colour gradient in the plot indicates the percentage of trees correct, it ranges from dark blue representing 0% correct to dark red representing 100% correct. The arrow marks the percentage of correct trees that would result from selecting at random from the 15 possibilities.

$c$ tends to infinity, the probability of MP recovering the correct tree tends to zero. The inconsistent zone can be seen in figure 2.6 in the box for MP with sequence length $c = 1000$, the portion shaded dark blue indicates that accuracy for this range of parameters is approaching 0%. CMP is consistent for the whole range of parameters.

With the two-state data UP was the most accurate of the methods tested, over all sequence lengths $c$ =100, 200, 500, and 1000. However, if UP was applied to data generated from a tree without equal rates, it would be expected to less accurate, this is because the UP method is not robust to violation of the molecular clock assumption [15, 46].

On the sequences generated for these simulations, ML is less powerful than most of the other methods, that is, it requires longer sequence lengths to attain the same degree of accuracy. For two-state sequences of length $c = 100$, ML is the least accurate of all the methods tested, however with four-state data it is only slightly ($\sim$ 2%) less accurate than NJ and UP. With the longest sequences tested ($c = 1000$) ML performs slightly worse than NJ and UP on two-state data, and slightly better

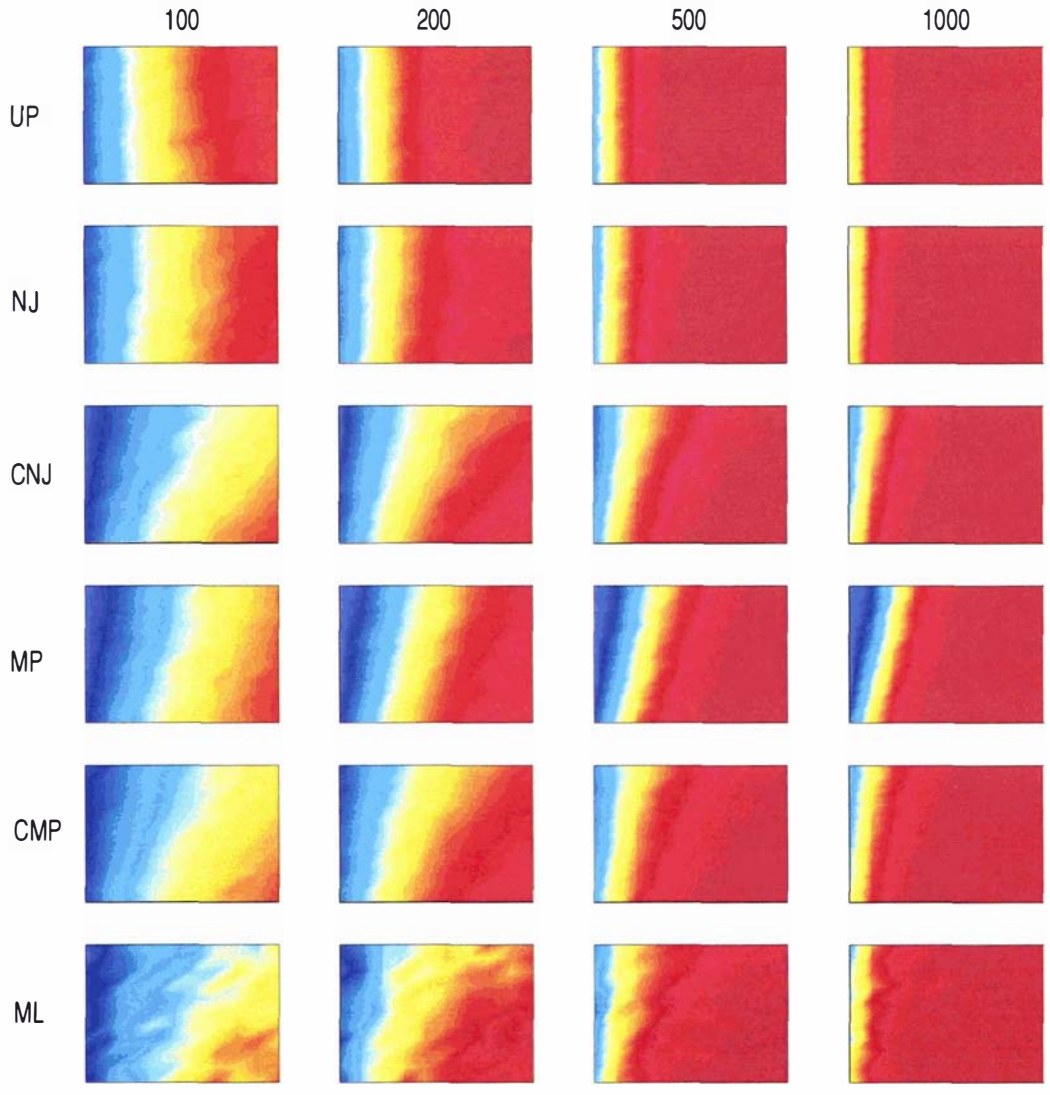


Figure 2.6: Accuracy with two-state data. These graphs show the different frequencies at which the trees were correctly recovered. Each box in the graph represents the complete parameter range for a specific method and sequence length. The horizontal axes show the length of the internal edges (0.01–0.10) and the vertical axes the length of the outgroup edge (0.20–0.40). Each row of boxes corresponds to a different tree estimation method. Each column represents a different sequence length, increasing from left to right. The shading within boxes indicates the percentage of trees correct, it ranges from dark blue representing 0% correct to dark red representing 100% correct.

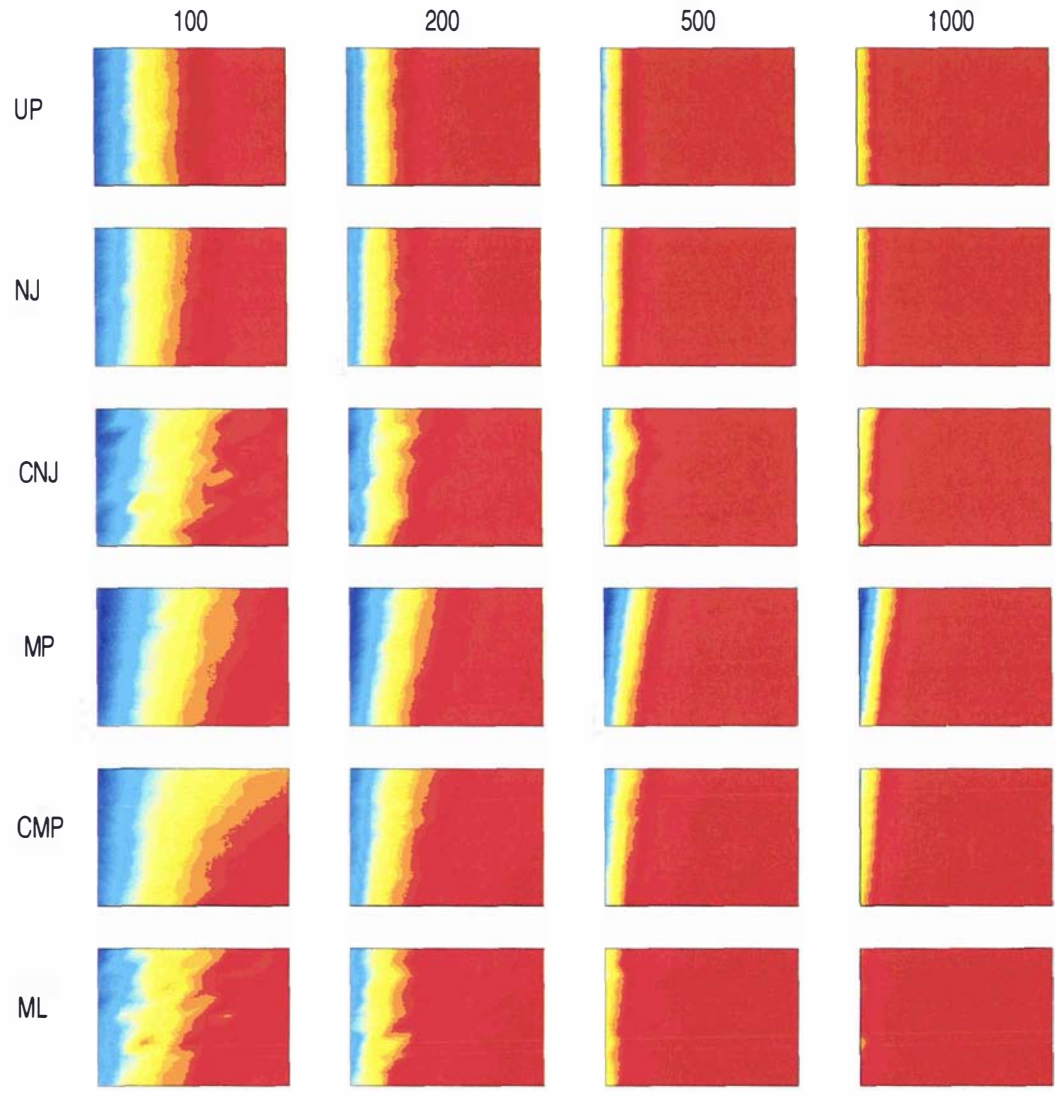


Figure 2.7: Accuracy with four-state data. These graphs show the different frequencies at which the trees were correctly recovered. Each box in the graph represents the complete parameter range for a specific method and sequence length. The horizontal axes show the length of the internal edges (0.01–0.10) and the vertical axes the length of the outgroup edge (0.20–0.40). Each row of boxes corresponds to a different tree estimation method. Each column represents a different sequence length, increasing from left to right. The shading within boxes indicates the percentage of trees correct, it ranges from dark blue representing $0\%$ correct to dark red representing $100\%$ correct.

on four-state data.

While all of the methods have lower accuracy with short internal edges, the accuracy of UP and NJ appears unaffected by the length of the outgroup edge. This is expected for UP because it is only influenced by the smallest distances when joining taxa. NJ joins the pair of taxa with the smallest *net divergence*. This quantity is changed by a common amount for each pair of taxa, when the length of an external edge is changed (see appendix 2.6 for proof). Thus, it is unsurprising that the length of the edge to the single outgroup has little effect on accuracy. In contrast, MP, CMP, ML and CNJ are less accurate in the regions where the outgroup edge is long.

### Correcting for multiple changes

For this data, when NJ is applied to distances that have been corrected for multiple changes there is a lower probability of recovering the correct tree than without using a correction. In the two-state case we compared the standard correction with the reduced bias correction of Tajima [98]. For sequence lengths of 100, over the whole range of parameters, NJ recovered 60.9% of trees correctly, with the reduced bias correction it recovered 51.6%, and with the standard correction only 46.4%.

The better performance of the methods on uncorrected distances could be due in part to the following two effects. The first is amplification of sampling error, the variance of the uncorrected estimate is less than the variance of the corrected estimate by a factor proportional to $e^d$, where $d$ is the corrected distance [98]. Secondly, the standard correction has a bias towards over-correcting; it tends to make the corrected distances too large. The reduced bias correction performs better than the standard distance correction although, for this data, it is not as good as no correction at all. As well as reducing the bias, this estimate also has a lower variance than the standard correction ; it is not clear whether the reduced bias or the lower variance is the major contributing factor in its improved performance over the standard distance correction.

These observations are consistent with other simulation studies [83, 91], where it was found that corrections for multiple substitutions are only helpful for recovering trees with unequal rates of change along edges. Kumar et al. [57], in the manual for their phylogeny software package MEGA, give guidelines for when it is beneficial to use a correction. They state that,

> "When the rate of nucleotide substitution is the same for all evolutionary lineages and the number of nucleotides used is relatively small, the $p$ [uncorrected] or Jukes-Cantor distance [standard correction] seems to give a correct tree more often than the Kimura distance even if there is some extent of transition/transversion bias (Schöniger and von Haeseler 1993; Tajima and Takezaki, 1994). When the substitution rate varies with evolutionary lineage, however, this is not the case."

The results of our analysis support their statement, trees were more accurately recovered using the uncorrected ($p$) distances than the Kimura (K2P) distances. Furthermore, in the two-state case, neighbor-joining was more accurate when applied to uncorrected ($p$) distances (NJ) than when applied to distances with the standard correction (CNJ).

Comparing MP with CMP we see that for short sequences ($c = 100$ or 200) accuracy is slightly worse when the Hadamard based correction is used. However, for longer sequences ($c = 500$ or 1,000) CMP becomes comparatively more accurate, because it is consistent whereas MP is not.

### Differences in two-state and four-state results

There were several differences between the two-state and four-state simulations. (Note that the expected number of changes along an edge was the same in each case, so in this sense it is valid to compare the results.) All methods were more accurate on the four-state data. Also, the zone in which MP is inconsistent is smaller. The performance of each of the methods, averaged over the range of parameters $x$ and

| | $c = 100$ | | $c = 1000$ | |
|---|---|---|---|---|
| Method | Two states | Four states | Two states | Four states |
| NJ | 60.9% | 72.3% | 94.0% | 96.7% |
| CNJ | 46.4% | 65.6% | 89.7% | 95.4% |
| UP | 67.0% | 72.8% | 94.7% | 95.9% |
| MP | 48.4% | 62.5% | 80.0% | 89.4% |
| CMP | 47.3% | 62.2% | 90.2% | 94.7% |
| ML | 44.6% | 70.0% | 92.6% | 98.0% |

Table 2.3: Summary of the accuracy of the methods with two-state and four-state data.

$y$, for $c = 100$ and $c = 1000$, is summarised in table 2.3. The difference in accuracy between NJ and CNJ for $c = 100$ was reduced from 14.5% to 6.7%, as the number of character states goes from two to four. The method with the largest difference in accuracy is ML, which for $c = 100$ is 25.4% more accurate on four-state data than two-state data. The reason that all the methods do better with four states than two is presumably the decreased frequency of parallel changes.

### Split decomposition

For the four-state model, the study was extended to include the method of split decomposition [5]. This method, as implemented in *SplitsTree* [49], does not force the data to fit a tree, but instead produces a set of weakly compatible splits. For this reason, we cannot make a direct comparison between the accuracy of *SplitsTree* and the other methods.

Split decomposition computes the isolation index of each split $A|B$ as follows:

$$\alpha_{A\,B} = \frac{1}{2} \min_{i,j \in A, k,l \in B} (\max\{d_{ij} + d_{kl}, d_{ik} + d_{jl}, d_{il} + d_{jk}\} - d_{ij} - d_{kl}),$$

where $d_{xy}$ is the (in these simulations, uncorrected) distance between taxa $x$ and $y$. Only those splits with non-negative isolation indexes are displayed in the split decomposition graph.

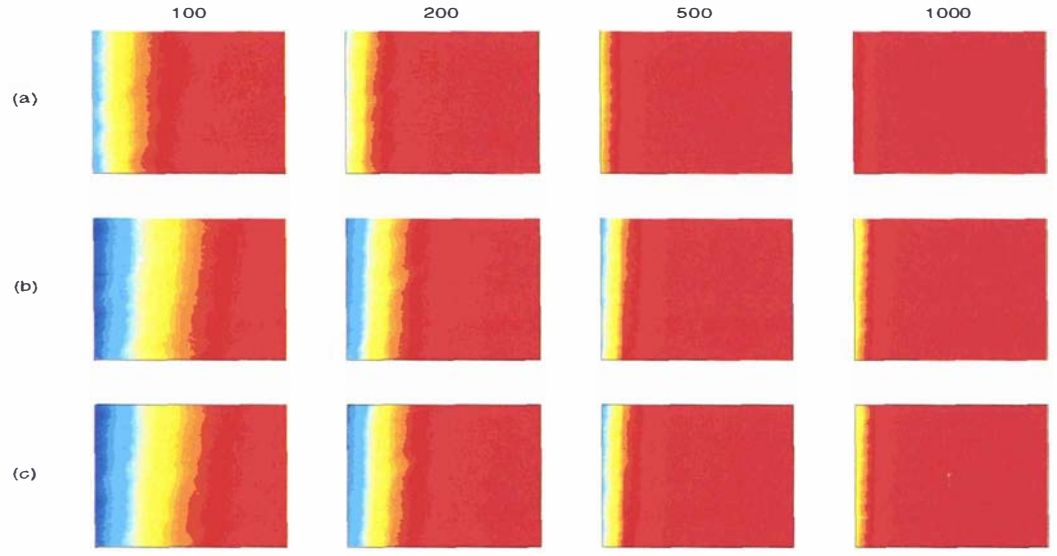


Figure 2.8: Accuracy of split decomposition with four-state data. As with the figures 2.6 and 2.7 each box in the graph represents the entire parameter range for a given method and sequence length. The horizontal axes show the length of the internal edges (0.01–0.10) and the vertical axes the length of the outgroup edge (0.20–0.40). Each column represents a different sequence length, increasing from left to right. The shading within boxes indicates the percentage of trees correct, it ranges from dark blue representing $0\%$ correct to dark red representing $100\%$ correct. Row (a) displays how often both of the correct splits had positive isolation indexes. Row (b) displays how often the tree formed from the pair of compatible splits with the largest sum of isolation indices was correct. Row (c) displays how often the greedy combination of two compatible splits gave the correct tree.

| | $c = 100$ | $c = 1000$ |
|---|---|---|
| Presence | 83.60% | 98.40% |
| Best Combination | 68.67% | 95.88% |
| Greedy Combination | 69.06% | 95.88% |

Table 2.4: Accuracy of split decomposition with four-state data. The first row shows the percentage of times the correct splits had positive isolation indices. The second row records how often the pair of splits with largest sum of isolation indices gave the correct tree, and the third row record how often the greedy combination of splits gave the correct tree.

The internal edges of the generating tree used in our simulations correspond to the the two non-trivial splits 12|345 and 125|34. The first measure of performance for ST records a success whenever both these splits have positive isolation index, the proportion of successes is shown in row (a) of figure 2.8. It is possible to have as many as five non-trivial splits with positive isolation index [4]. This means that the two "true" splits will have a much greater opportunity to be included within a split decomposition graph than in a tree (formed by other methods) where there can be at most two non-trivial splits represented. As expected, ST does very well according to this criterion.

In order to provide a comparison to the methods previously discussed, I chose two methods of deriving trees from the split decomposition, these were suggested in Bandelt and Dress (1992) [5]. The first method picks the pair of compatible splits with the largest sum of isolation indices. The second is a greedy algorithm where the split with the largest isolation index is chosen first, followed by the next largest split that is compatible with it. In each case, when the two selected splits are 12|345 and 34|125 a success is recorded. The results from these two approaches are displayed in rows (b) and (c) of figure 2.8. The two methods have very similar accuracy. Both methods are more accurate than MP, CMP and CNJ and slightly less assurate than NJ, UP and ML. Table 2.4 records the percentage accuracy of the three measures of performance for split decomposition, for $c = 100$ and $c = 1000$.

### Asymmetry of internal edges

The model tree on which these simulations were based is symmetrical. For the two-state model, the simulation was also run with skewed trees where $q_{12\,345}$ and $q_{34|125}$ differ, and the internal edges are adjusted so that the molecular-clock still holds. This did not greatly affect the patterns of accuracy seen in figure 2.6 (results not shown). It seems that the length of the edge that the outgroup joins to $(q_{12|345} + q_{34|125})$ has a greater impact on the accuracy than the amount of asymmetry $(|q_{12|345} - q_{34|125}|)$. However, if the outgroup joins close to one end of the internal edge of the ingroup tree it will be more prone to error than if it joins in the middle.

### Summary

The existence of a molecular clock has an impact on the relative accuracy of tree estimation methods. It does not guarantee consistency for MP, but NJ and UP are consistent on equal rate trees without requiring corrected distances [14]. For the trees used in this simulation the distance based methods NJ and UP outperform MP, CMP and ML. NJ was more accurate than its corrected counterpart CNJ over the entire parameter space tested. However, correcting parsimony was helpful within the range of parameters for which MP is inconsistent if the sequences were long enough ($c$ >200–500).

## 2.4.2 Methods can be Consistent but Misleading

It is well known that MP can be inconsistent with unequal rates of evolution. It can be seen from figures 2.6 and 2.7 that it can also be inconsistent for trees on five taxa with a molecular clock. This inconsistent zone occurs when the internal branches are short and the outgroup branch is long. For the two-state case on the

tree of figure 2.3, this occurs when

$$y > -\frac{1}{2}\log\left[\frac{1}{1-A^2}(2Ae^{-2x} - A(1+A^2)e^{2x})\right], \tag{2.2}$$

with $A = e^{-0.2}$ and $x$ and $y$ as shown in figure 2.3. (See appendix 2.6 for derivation.)

If the parameters are outside the inconsistent zone for MP, then as the sequence length $c \to \infty$, the probability that MP recovers the correct tree tends to one. However, this does not guarantee that the correct tree will be recovered with greater probability than each of the other trees, for a fixed sequence length. Indeed, for some parameter choices outside, but close to the boundary of, the inconsistent zone, we discover that a group of specific incorrect trees are each selected with greater frequency than the generating tree. This effect diminishes as the parameters move away from the boundary.

Consider the tree in figure 2.3 with $x = 0.016$ and $y = 0.3$. These parameters place the tree just inside the consistent zone for the MP method, that is, the expected parsimony score for the generating tree (((1,2),(3,4)),5) is less than for all other possible trees. However, consistency alone does not indicate how long sequences should be in order to recover the correct tree with high frequency. To estimate the relationship between the length of the sequences and the accuracy of the methods (NJ, CNJ, UP, ML, MP, CMP), 100,000 sequence samples were generated for each length $c \in \{10^2, 10^3, 10^4\}$. The results of this experiment for the two-state model are shown in table 2.5. As expected, with short sequences, an incorrect tree is often selected. What was surprising was that for MP, four incorrect trees were each selected with greater frequency than the correct tree. These were the four trees where the ingroup phylogeny is correct, but the outgroup (taxon 5) is incorrectly joined to one of the external edges. Although it is not apparent with these parameter values, we later show that the other methods also suffer from the same problem, but to a lesser degree.

For MP the simulations were extended to include more values of $c$ within the

range 100–100,000, the results are displayed in figure 2.9. With $c = 100$, MP selected the correct tree (((1,2),(3,4)),5) in only $11.3 \pm 0.3\%$ of trials, and each of the four trees where the outgroup is incorrectly joined to a pendant edge in $14.2 \pm 0.3\%$ of the trials. The remaining ten trees were selected on average in $3.2 \pm 0.2\%$ of trials. With $c = 1000$, MP selected the correct tree in $14.6 \pm 0.4\%$ of trials, and each of the other four trees $21.1 \pm 0.6\%$. Thus, even though the expected parsimony length of the generating tree (((1,2),(3,4)),5) is shortest, each of the four other trees was selected with a significantly greater frequency. Over the range of sequence lengths $c$ =100–1000, the proportion of trials in which the four incorrectly rooted trees are chosen increases faster than for the correct tree. Not until $c = 5,000$ is the correct tree most frequently selected. Eventually, for sequences of length $c = 10^6$, only the correct tree is selected.

We now explain this counter-intuitive result for MP. MP uses the Fitch algorithm to calculate the number of changes required at each site on a given tree. In the five taxon case the total cost can be decomposed as follows:

Let $L(T)$ be the parsimony length of a tree, $T$.

$$\begin{aligned} L(T) &= \#S + \#C + 2\#I \\ &= (2c - 2\#K - \#S) - \#C \\ &= M - \#C, \end{aligned}$$

where,

K = constant sites,

S = singleton sites,

C = non-singleton sites compatible with the edge splits of $T$,

I = non-singleton sites incompatible with the edge splits of $T$, and

$c$ = sequence length.

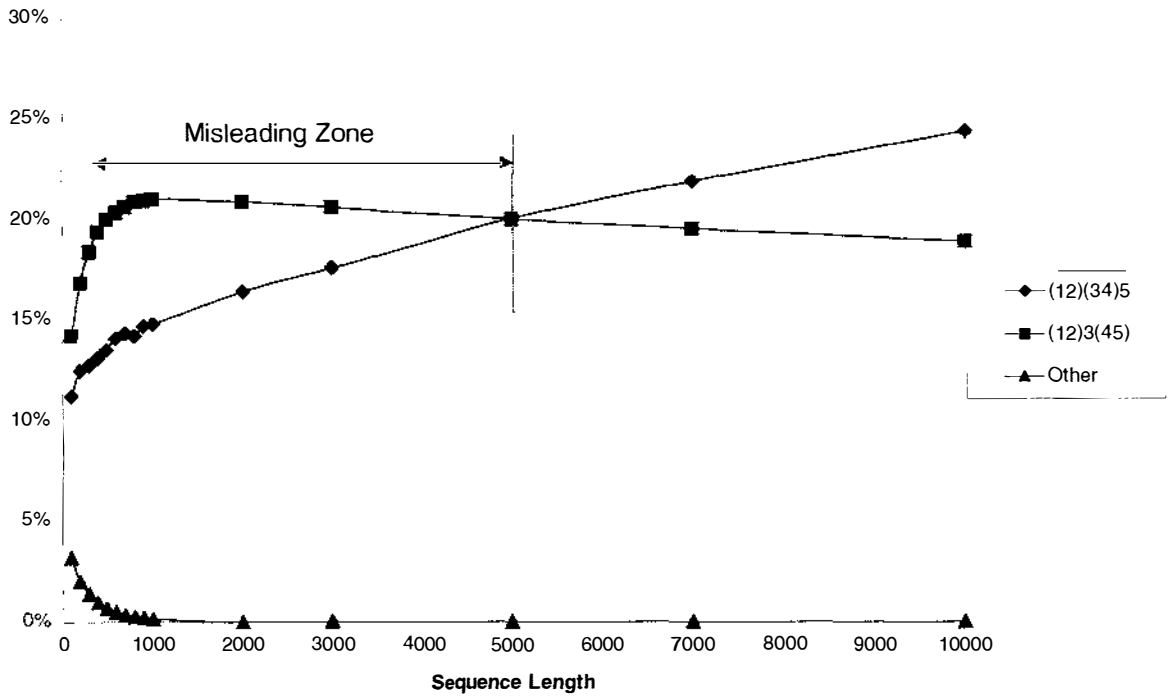


Figure 2.9: The Misleading Zone for MP. Each point on the graph represents 100,000 repetitions. The misleading zone extends to approximately $c = 5000$. The line for (12)(34)5 shows the frequency with which the correct tree was chosen. (12)3(45) is one of the four trees where the outgroup has incorrectly been joined to an external edge. "Other" is the average of the ten trees with an incorrect ingroup. The tree for this simulation has parameters $x = 0.016, y = 0.3$, see figure 2.3.

| Method ($c = 100$) | (1 tree) Correct Tree | (each of 4 trees) Outgroup Wrong | (each of 10 trees) All Other Trees |
|---|---|---|---|
| MP | 11.3% | 14.2% | 3.2% |
| CMP | 18.5% | 12.0% | 3.4% |
| NJ | 22.2% | 12.3% | 2.9% |
| CNJ | 13.8% | 13.8% | 3.1% |
| UP | 24.9% | 9.3% | 3.8% |
| ML | 16.2% | 14.0% | 2.8% |
| Method ($c = 1000$) | Correct Tree | Outgroup Wrong | All Other Trees |
| MP | 14.6% | 21.1% | 0.1% |
| CMP | 64.0% | 8.4% | 0.2% |
| NJ | 78.9% | 5.2% | 0.0% |
| CNJ | 59.2% | 9.9% | 0.1% |
| UP | 83.6% | 3.2% | 0.3% |
| ML | 74.6% | 6.3% | 0.0% |
| Method ($c = 10,000$) | Correct Tree | Outgroup Wrong | All Other Trees |
| MP | 24.6% | 18.9% | 0.0% |
| CMP | 100.0% | 0.0% | 0.0% |
| NJ | 100.0% | 0.0% | 0.0% |
| CNJ | 99.3% | 0.2% | 0.0% |
| UP | 100.0% | 0.0% | 0.0% |
| ML | 100.0% | 0.0% | 0.0% |

Table 2.5: The frequencies with which different trees were recovered. In all cases the standard error of these values was small ($< 1\%$) so they are not reported in the table. For the generating tree ($x = 0.016$, $y = 0.3$, see figure 2.3) all these methods are consistent.

$M = 2c - 2\#K - \#S$ is the same for each five taxon tree for a given data set.

$\mathbf{s}c$ represents the expected sequence spectrum. Let $\hat{\mathbf{s}}$ represent the observed sequence spectrum (selected from $s$), for each split $\alpha$ of the taxon set, $\hat{\mathbf{s}}_\alpha$ is the number of sites that induce the split $\alpha$. So, the correct tree, $T_0$, and four most frequently occurring incorrect trees, $T_1, T_2, T_3$ and $T_4$, have the following parsimony lengths:

$$L(T_0) = L(((1,2),(3,4)),5) = M - \hat{s}_{12|345} - \hat{s}_{34|125}$$

$$L(T_1) = L(((1,5),2),(3,4)) = M - \hat{s}_{15|234} - \hat{s}_{34|125}$$

$$L(T_2) = L(((2,5),1),(3,4)) = M - \hat{s}_{25|134} - \hat{s}_{34|125}$$

$$L(T_3) = L(((1,2),3),(4,5)) = M - \hat{s}_{12|345} - \hat{s}_{123|45}$$

$$L(T_4) = L(((1,2),4),(3,5)) = M - \hat{s}_{12|345} - \hat{s}_{124|35}$$

Now, due to the symmetry of our model tree, $s_{15|234} = s_{25|134} = s_{35|124} = s_{45|123}$ and $s_{12\,345} = s_{34\,125}$. This means that each of the four incorrect trees, $T_1, T_2, T_3$ and $T_4$ has the same expected length. Furthermore, at the boundary of inconsistency, these six $s_\alpha$ values are equal, hence each of the five trees $T_0, \ldots, T_4$ has equal expected length, and this is smaller than the expected length of each of the other ten possible trees.

However, for simulations with the parameters at the boundary of inconsistency ($x = 0.014825$ and $y = 0.3$), MP does *not* select each of the five competing trees with equal frequency. The parsimony length of the correct tree, $T_0$, is minimal only when both:

- $s_{12|345}$ is greater than both $s_{15|234}$ and $s_{25|134}$
- $s_{34|125}$ is greater than both $s_{35\,124}$ and $s_{45|123}$.

| $c$ | (1 tree) Correct Tree | (each of 4 trees) Outgroup Wrong | (each of 10 trees) All Other Trees |
|---|---|---|---|
| 100 | 10.4% | 14.0% | 3.4% |
| 1,000 | 11.3% | 21.7% | 0.2% |
| 10,000 | 11.1% | 22.2% | 0.0% |
| 100,000 | 11.1% | 22.2% | 0.0% |

Table 2.6: With the parameters set to $x = 0.014825$ and $y = 0.3$ (from equation 2.2), the model tree is at the boundary of consistency (to 5dp). The frequencies converge to $\frac{1}{9}$ and $\frac{2}{9}$ as the sequence length increases.

Each of the six $s_\alpha$ terms have equal expected value. Hence, $s_{12|345}$ is expected to be greater than both $s_{15\,234}$ and $s_{25\,134}$ in one third of the samples, and, independently, $s_{34\,125}$ is expected to be greater than both $s_{35|124}$ and $s_{45|123}$ in one third of the samples. Thus, we expect $L(T_0)$ to be minimal amongst the five trees in approximately one ninth of the samples. When the sequence length $c$ is small, one of the remaining ten trees could be minimal, but as $c$ is increased, the proportion of samples where $T_0$ is selected will tend to $\frac{1}{9}$, while the frequency of each of the other four trees will tend to $\frac{2}{9}$. Our simulations at the boundary of inconsistency supported these conclusions as can be seen in table 2.6.

We observe this effect changing as the parameters move off the boundary into the consistent zone. Consider altering the $x$ and $y$ parameters slightly so that the tree is consistently estimated by MP, for example, to the parameters used in table 2.3 and figure 2.9 ($x = 0.016$ and $y = 0.3$). The expected frequency of the correct splits are $s_{12|345} = s_{34|125} = 0.0172$ and the sampling variance is $V(s_{12|345}) = 0.01689/c$. The expected frequency of the four splits where taxon 5 joins to taxon 1,2,3 or 4 is $s_{15|234} = 0.0165$ and the sampling variance is $V(s_{15|234}) = 0.01621/c$ [104]. Although this tree is consistently estimated by MP, it is not until $c$ becomes large ($c > 5000$), and hence the sampling variance is small, that the correct tree is most frequently chosen. Figure 2.10 shows the expected frequencies of the non-trivial splits for the tree with $x = 0.016$, $y = 0.3$.

We investigated further to see if other tree-estimation methods suffered from a

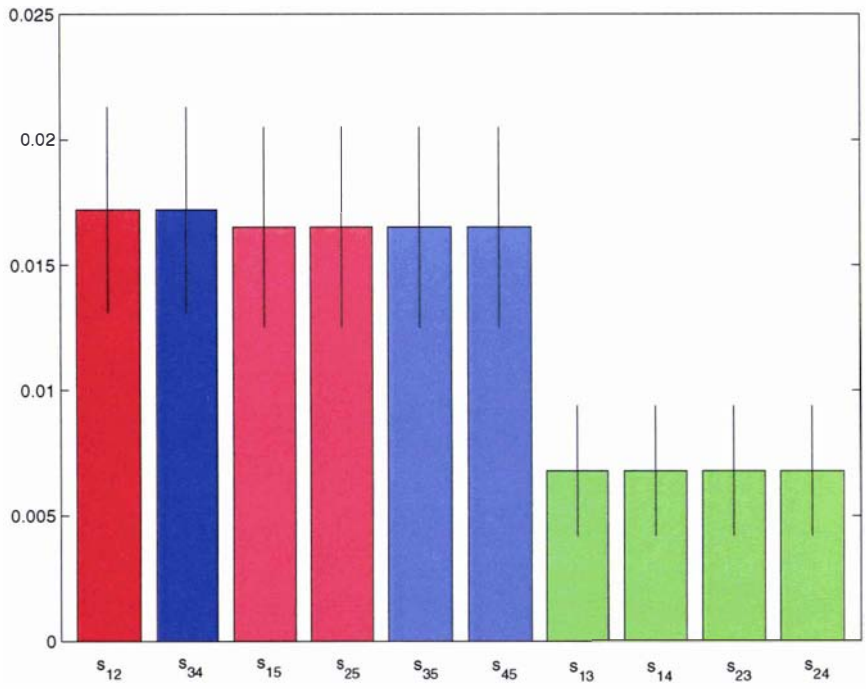


Figure 2.10: The expected frequencies of the non-trivial splits for the tree with $x = 0.016$, $y = 0.3$. Note that the splits $\alpha$ are labelled with only one half of the partition $A|B$. The two correct splits $12|345$ and $34|125$, have the largest expected values $s_{12}$ and $s_{34}$, shown in bright blue and red, but if either of the observed frequencies $\hat{s}_{15}$ or $\hat{s}_{25}$ are larger than $\hat{s}_{12}$, or either of $\hat{s}_{35}$ or $\hat{s}_{45}$ are larger than $\hat{s}_{34}$ then an incorrect topology will be chosen. The line crossing each bar shows $\pm$ one standard deviation for the observed frequency of each split for $c = 1000$. As $c$ gets larger the standard deviation becomes smaller, and the probability that $\hat{s}_{12}$ and $\hat{s}_{34}$ will be the largest two splits in the observed sequence spectrum tends to 1.

"misleading" effect, similar to the one illustrated for MP in figure 2.9, with more extreme values of the parameters. We kept the sample size $c = 100$ and the edge parameters as they are shown in figure 2.3 with $y = 0.3$, the value of $x$, the length of the internal edges, was increased from zero until the frequency of the correct tree was higher than the frequency of each of the four trees with the incorrectly joined outgroup. For NJ and UP the value of $x$ only had to be increased to $\simeq 0.005$ for the frequencies to become equal. The effect lasted longer for other methods, CP $x \simeq 0.006$, ML $x \simeq 0.008$, CNJ $x \simeq 0.016$ and for MP $x \simeq 0.020$. So, all the methods tested here suffer from this problem but to a lesser extent than MP.

To gain some insight into why all the methods have a "misleading" zone, consider the generating tree $T_0$ in the case where $y = 0.3$ as $x \to 0$. At $x = 0$ this is the star tree, where the external edges leading to the taxa 1, 2, 3, 4 have length 0.1 and the edge leading to the outgroup taxon 5 has length 0.3. We classify the fifteen possible trees on five taxa into two classes. Twelve where the outgroup is paired with another taxon (type 1), and three where the outgroup is not part of a neighbouring pair (type 2), the correct tree $T_0$ is of type 2. Simulations at this boundary tree reveal that all of the methods studied here have a bias towards choosing trees of type 1, where the outgroup (taxon 5) is part of a neighbouring pair with another taxa, results are shown in table 2.7. A continuity argument suggests that on increasing $x$ slightly from zero there will still be a bias for selecting an incorrect (type 1) tree.

The existence of these "misleading" zones, where the tree estimate is consistent but other trees are selected more frequently for bounded sequence lengths, does not appear to have been previously reported. However, other unexpected tree estimation effects have been discussed in the literature. These include a simulation study by Yang (1997) [110] showing that for some trees ML is more accurate assuming an incorrect (simpler) model of evolution than that used to generate the data. Bruno and Halpern (1999) [11] suggested that this is due to the simple model being biased towards the tree that in this instance happens to be correct. They report

| Method | $c = 100$ | $c = 1000$ |
|---|---|---|
| NJ | 11.65% (0.15) | 11.17% (0.22) |
| UP | 11.04% (0.12) | 10.36% (0.11) |
| MP | 3.59% (0.10) | 0.00% (0.00) |
| ML | 9.06% (0.71) | 10.32% (0.62) |

Table 2.7: Data was simulated along a star tree where the edges to taxa 1, 2, 3, and 4 had length 0.1, and the edge to taxon 5 had length 0.3. The table records the percentage of times that a tree is estimated where the outgroup, taxon 5, does *not* form a neighbouring pair with any other taxa. There are three out of fifteen such trees. However, these are chosen significantly less frequently than $\frac{3}{15} = 20\%$ of the time. Each entry represents $10 \times 10,000$ repetitions, with the exception of ML for which $10 \times 1,000$ repetitions were performed. The number in brackets is the standard error.

that for other trees the true model (i.e. the one used to generate the sequences) is superior to the simple model. Posada and Crandall (2001) [75] give an example of fortuitously biased simple models leading to, what they suggest are, better tree estimates in retroviruses. Kim (1993) [52] in a simulation study of UP, MP and NJ states that "Both MP and NJ methods estimate certain kinds of topologies more frequently than others and in a more marked way than does the UP method." A topological bias for some methods was also reported by Charleston (1994) [14, 15].

A recent paper by Kim (2000) [53] suggests a geometric framework for studying the interaction between the tree and tree estimation methods in order to better understand the above effects. Following on from Efron et al. (1996) [26], Kim (2000) considers the space of all possible data sets. Each tree estimation method divides this space into different regions, each region representing the tree that the method estimates on those data points. In figure 12 of his paper, Kim [53] outlines a hypothetical scenario where an inconsistent method becomes more accurately estimated over a finite range of sequence lengths. The "misleading" effect that we describe is the converse to Kim's example. Here a consistent method is progressively less accurately estimated over a finite range of sequence lengths.

### 2.4.3 Classes of Error in Placing the Root

Given that a tree has been recovered incorrectly, it is informative to categorise different errors that can occur. We divide the incorrect trees into two classes: (ingroup correct) trees where it is only the outgroup (taxon 5) that has been placed incorrectly; and (ingroup incorrect) trees where the structure of the ingroup (taxa 1,2,3,4), in the five-taxon tree, is incorrect. This second class of trees was re-estimated on the ingroup alone. This second class of trees is divided into two sub-classes: those where the ingroup is still incorrect; and those where the ingroup is correct. In this last subclass the inclusion of the outgroup has had a confounding effect, that is, adding the outgroup causes the ingroup tree to be incorrectly recovered. This categorisation of errors is summarised in figure 2.11, where the percentage of trials resulting in each category is reported.

It can be seen from figure 2.11 that the most common cause of error is that the method has joined the outgroup to the wrong edge of the ingroup tree. Of the fourteen incorrect trees on five taxa, four of them have the correct ingroup structure but the outgroup is wrongly placed. The proportion of these four trees greatly outweighs that of the ten trees with incorrect ingroup structure.

The methods vary in the frequency with which they are confounded by the inclusion of the outgroup. With NJ, ML and MP the addition of an outgroup can decrease the accuracy with which the ingroup is recovered, but with UP this is very rarely observed. In the simulations with two-state characters NJ was the most likely method to be confounded. This may be because the NJ criterion is dependent on all pairwise distances, whereas UP identifies only the smallest entry in the distance matrix and joins the corresponding pair. Thus for UP, the inclusion of the outgroup has little effect on how it constructs the rest of the tree.

Another question is whether the inclusion of an outgroup can ever *improve* the accuracy of tree estimation. To test this we constructed trees separately both with and without the outgroup. We found it was rare ($< 1\%$) that any of these methods

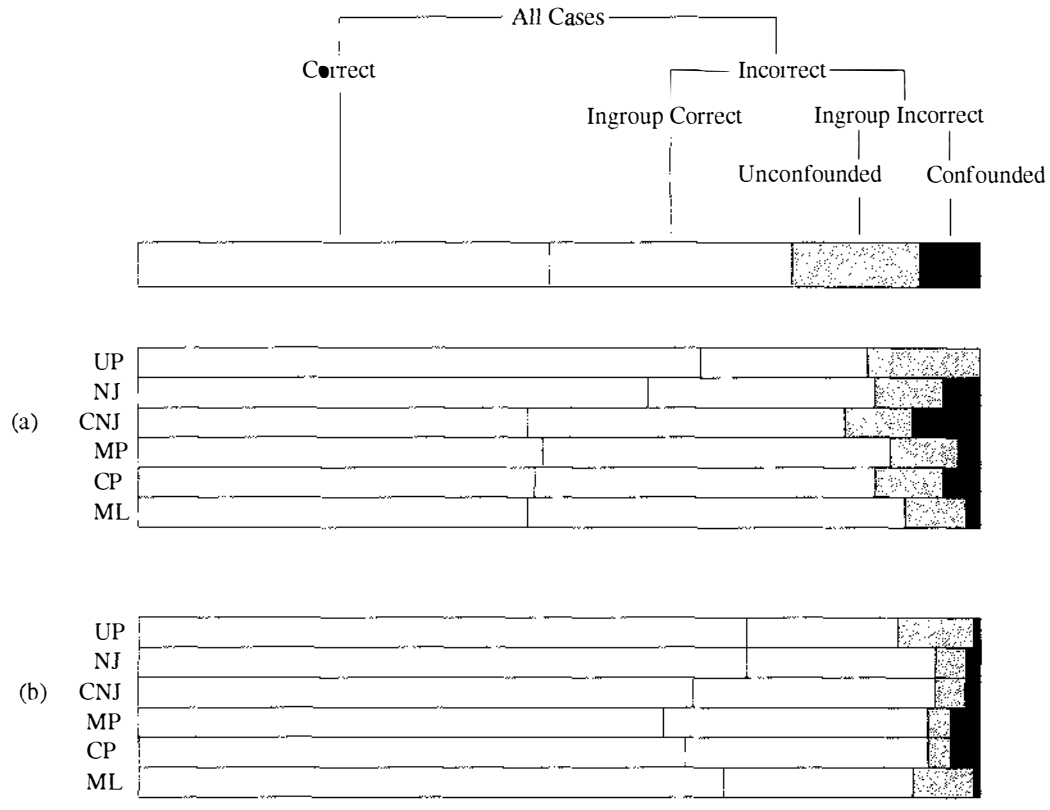


Figure 2.11: Frequencies of the different types of error. The results are shown for sequences of length $c = 100$ with the two-state model (a), and four-state K2P model (b). The top bar displays the shading scale for the different types of error. "Ingroup correct" are those trees where the only error is the misplacement of the outgroup. The "ingroup incorrect" trees are those where the internal structure of the tree is wrong. Trees are said to be "confounded" if the four taxon ingroup tree was correct and only became incorrect when the outgroup was added. A tree is "unconfounded" if it was wrong regardless of including the outgroup. The bars for NJ, CNJ, UP, MP and CMP represent averages from 10,000 repetitions of each the parameter values described in table 2.1. The results for ML are averages from 1000 repetitions of the parameter values.

would produce an incorrect ingroup tree, but find the correct five-taxon tree. It is more frequently observed that the inclusion of an outgroup disrupts the ingroup tree. For this reason, it is good practice to construct trees both with and without the outgroup. If the topology of the ingroup changes when the outgroup is included in the analysis, it is likely that the tree estimated without the outgroup is more accurate.

Further work that could be done in this area is to check the impact of including an outgroup in trees with greater numbers of taxa. Also, it would be interesting to investigate how the accuracy of outgroup placement is influenced by using a larger number of outgroup taxa.

## 2.5 Discussion

Several interesting effects have been reported in this chapter and are relevant to practical examples. Four effects are discussed in greater detail: the existence of a zone of inconsistency for MP with equal rates; the detrimental effect of correcting for multiple changes with short sequences; the newly reported zone where methods are consistent but misleading; and finally, the difficulty of getting an outgroup to correctly join with a short edge of an ingroup tree.

The simulation study illustrates the observation of Hendy and Penny (1989) [43] that MP can be inconsistent with equal rates if there is a combination of short and long edges within the tree. Correcting parsimony for multiple changes (CMP) via Hadamard conjugation makes it consistent. In the five-taxon case presented here MP is consistent when applied to each of the four quartets of the tree, but inconsistent when applied to the tree as a whole.

It was seen that correcting for multiple changes is not always beneficial. Clearly if the generating tree has significantly different rates of change along its edges then correction is necessary for methods to remain consistent. However, if the generating tree obeys a molecular clock then corrections for multiple changes will both bias and

increase the variance of the distances, without improving the chances of recovering the correct tree. Our results suggest that for short sequence lengths ($< 500$ sites, not including sites that are invariable for biochemical reasons) where the assumption of equal rates cannot be rejected, one should test the effect of omitting corrections for multiple changes.

We discovered a misleading zone where, although methods are consistent, for bounded sequence lengths the correct tree was recovered less frequently than some other individual incorrect trees. This effect was most pronounced for MP where for the correct tree to be chosen it had to "win" two independent competitions between splits ($s_\alpha$), whereas the incorrectly rooted topologies only required one strong supporting pattern to be chosen.

Finally, rooting a tree by joining a distant outgroup to a small internal edge will frequently cause problems. Usually only the placement of the outgroup will be incorrect, not the internal structure of the tree. However, some methods are susceptible to being confounded by the inclusion of an outgroup. This suggests that it is good policy to make a phylogeny both with and without an outgroup and compare the results. In this study, when the rooted and unrooted ingroup structures differed, then the unrooted tree was more frequently correct. I expect that this would also be true for trees with greater numbers of taxa. Although this study has been specifically looking at five-taxon trees, it is expected that the problems raised are pertinent to larger trees, as five-taxon subtrees of this type will be embedded within larger trees.

Error free reconstruction of phylogenetic trees is impossible, however, the more that is known about the performance of different methods under a range of conditions the better the chance of detecting errors. While there is no universally successful method, some methods will be better than others for a given data set. For a small number of taxa, if a data set appears to be clock-like then an uncorrected distance based method such as UP or NJ may be more accurate than MP or ML.

## 2.6 Appendix

### Derivation of equation 2.2

Recall that for the tree shown in figure 2.3 $s_{12} = s_{34}$ and $s_{15} = s_{25} = s_{35} = s_{45}$. Maximum parsimony will be inconsistent if and only if $s_{15} - s_{12} > 0$ which implies

$$\sum_{\alpha}(h_{\alpha\cdot 234} - h_{\alpha\cdot 12})r_{\alpha} > 0$$

or, equivalently

$$r_1 + r_{23} + r_{24} + r_{134} > r_{12} + r_3 + r_4 + r_{1234} \tag{2.3}$$

where Hadamard entry $h_{\alpha\cdot\beta} = (-1)^{\alpha\cap\beta}$ and $\mathbf{r} = \exp(H\mathbf{s})$ (Note, see [45] for a description of how the $\mathbf{s}$ and $\mathbf{r}$ vectors are indexed.)

Let $A = e^{-0.2}$, $B = e^{-2x}$ and $C = e^{-2y}$ with $x$ and $y$ as in figure 2.3, $0 \leq x \leq 0.1$ and $y > x + 0.1$.

$$r_1 = r_3 = r_4 = e^{-2(0.1+x+y)} = ABC$$

$$r_{23} = r_{24} = e^{-2(0.2+2x)} = A^2B^2$$

$$r_{134} = e^{-2(0.3+x+y)} = A^3BC$$

$$r_{12} = e^{-2(0.2)} = A^2$$

$$r_{1234} = e^{-2(0.4)} = A^4$$

Substituting these values for $r_{\alpha}$ into 2.3 we find that MP is inconsistent if,

$$2A^2B^2 + A^3BC > A^2 + ABC + A^4,$$

or, rearranging

$$C < \frac{A(1 + A^2 - 2B^2)}{A^2B - B}$$

$$C < \frac{A}{1 - A^2}(2B - B^{-1} - A^2B^{-1})$$

$$y > -\frac{1}{2}\ln\left[\frac{A}{1-A^2}(2e^{-2x} - (1+A^2)e^{2x})\right]$$

∎

**NJ is unaffected by the length of the outgroup edge**

**Proposition**: *The* **net divergence** $S_{ij}$ *is changed by a common amount for all* $i, j$ *when the distance from one taxon,* $a \in X$*, to each other taxon is changed by a common amount.*

**Proof**: Let $D$ be a distance matrix on a set $X$ of $n$ taxa. The net divergence is $S_{ij} = (n-2)D_{ij} - R_i - R_j$, where $R_i$ is the sum of the distances in row $i$.

Form a new matrix $D^*$ by increasing all the distances to the taxon $a \in X$ by a constant amount $x$.

Now, for $i \neq a, j \neq a$

$$\begin{aligned} S^*_{ij} &= (n-2)D_{ij} - (R_i + x) - (R_j + x) \\ &= S_{ij} - 2x \end{aligned}$$

and, for $i = a, j \neq a$

$$\begin{aligned} S^*_{aj} &= (n-2)(D_{aj} + x) - (R_a + (N-1)x) - (R_j + x) \\ &= S_{ij} - 2x \end{aligned}$$

As $S_{ij}$ is symmetric this proves our proposition. ∎

# Chapter 3

# Detecting evolution in Adélie Penguins: A network based approach

## 3.1 Introduction

This chapter presents work initiated in collaboration with Massey ecologists David Lambert and Peter Ritchie [58]. They have collected a large sample of sequences from HyperVariable Region I (HVRI) of the Adélie penguin (*Pygoscelis adeliae*) mitochondrial genome, the data comprises both ancient and modern mitochondrial (mt) DNA. The ancient mtDNA has been recovered from bones preserved in ornithogenic soils [42]; many of these bones have been carbon dated. The Adélie data set provides a unique opportunity to observe HVRI in snapshots through time.

The central questions motivating my involvement in this project are: Has measurable evolution taken place in Adélie penguins over the time scale represented, and furthermore, at what rate is HVRI evolving? Another important aspect is to judge how best the ongoing collection of modern and ancient samples should be carried out to help answer the above questions.

In the following section I give some background on the data collection and

some relevant properties of HVRI. In section 3.3 I discuss the simulation study of the birth/death process in Adélie penguin populations, which aimed to estimate statistical properties of the coalescent process. In section 3.4 a discovery curve is fit to the data in order to estimate the proportion of haplotypes that have been sampled. Section 3.5 contains preliminary phylogenetic analysis of the data, which reveals why standard tree estimation techniques are unsatisfactory in this case. In section 3.6 a median network [3, 6] is developed for the data, and the subgroups defined in the median network are tested for correlation to geographical location. Lastly, in section 3.7, a method is developed that uses the network of modern and ancient samples to estimate the rate of evolution for HVRI, this method is applied to the data set. A randomisation test is conducted to measure the extent of the bias in the rate estimation method.

The work in this chapter has been influenced by the help of a number of people who have offered both expertise and financial support. In particular I would like to acknowledge: Andreas Dress and Katharina Huber for their advice on using median networks; Bill Martin for interesting discussion; Allen Rodrigo and Alexei Drummond for their suggestions on how to validate the median network rate estimation method and other statistical advice. Thanks go to Andreas Dress, Vincent Moulton and Allen Rodrigo for supporting me with their various grants and hosting me at their institutions.

## 3.2 Background

Adélie penguins live in large colonies along the coast of Antarctica [109]. Over the Antarctic summer they return to the coastal rock shelves in order to breed. During this time the Adélie penguin chicks, in particular, are subject to high levels of mortality. Their remains are usually left in their nesting areas, so over many thousands of years deposits of penguin bones and guano have built up, forming ornithogenic soils down to $\sim$1 meter. Once the bones are beneath the surface, they

stay at a roughly constant sub-zero temperature; cold, dry conditions like those found in the Antarctic have been found to preserve DNA [59].

Using ancient DNA technology [48, 71] Peter Ritchie was able to extract high quality mtDNA from bones carbon dated as old as 6082 years [82]. By digging down through abandoned penguin colonies it was possible for Ritchie and team to compile a data set of serially preserved DNA samples. In addition they collected and sequenced a large number of blood samples from living Adélies. This data set provides a unique opportunity to examine evolution in action.

As of May 2001 the data set consisted of sequences from 322 modern (living) Adélies, and 79 ancient bones whose carbon dates range back to 6082 years. The amount of sequence aligned over all 404 samples is 353bp. The sequences come from the control region of the mitochondrial genome, in particular the section known as HyperVariable Region I (HVRI).

Mitochondrial DNA is typically passed down through the female line and is not thought to be subject to recombination [9]. This implies that it should, in theory, be possible to use mtDNA to reconstruct a tree representing the mitochondrial genealogy of samples within a species. For example, HVRI has been previously used to study patterns of migration in human populations [6, 8, 106]. Rates in HVRI are thought to be highly variable across sites, and transitions are far more common than transversions [65]. These factors mean that the data set is likely to contain a lot of homoplasy (parallel changes and reversals), this makes reliable tree construction difficult.

In most vertebrate species HVRI evolves quickly compared to other parts of the mt genome [87, 1]. In a study of snowgeese Quinn [78] estimated that HVRI evolves 10.4 times faster than the average rate for the mt genome. The currently accepted rate for HVRI, based on phylogenetic information in combination with the fossil record, is 0.208 substitutions per site per million years (s/s/Myr) [78]. Based on this rate of 0.208 s/s/Myr we would expect there to have been only 0.44 changes over the 353bp aligned region between a 6000yr old sample and any one of

its currently living direct descendants. One aim of this study is to test if a direct estimate of the rate based on the ancient DNA corresponds to the rate suggested by the fossil record and phylogenetic evidence.

When David Lambert and Peter Ritchie began this project they hoped to be able to identify direct descendents of ancient samples. Given such ancestor descendent pairs, it would be possible to count the number of differences in their DNA sequences, divide this by the carbon date, and thus estimate the rate of mutation. These estimates could be combined for each ancestor-descendant pair by taking an average. There is merit in the simplicity of this idea, but unfortunately there are obstacles. Firstly, there is no guarantee that a particular ancient sequence will come from a bird with any living descendants, alternatively, an ancient sequence may have existing descendants that are not represented in the modern sample. Even given the existence of ancestor-descendent pairs within the sample, it will not be trivial to identify them on the basis of sequence similarity over the HVRI region alone.

To focus on the sampling problem imagine that all of the ancient samples are the same age. Then for a given living bird the probability of sampling its direct ancestor is $\frac{N_s}{N}$ where $N_s$ is the size of the ancient sample, for the Adélie data about 80, and $N$ is the female population size, say 10,000 for an average colony. For this example the probability of sampling a direct ancestor-descendent pair is 0.008. (Note that our ancient samples have actually been collected from a number of different colonies.) This highlights how unlikely it is that we have sampled any direct ancestor-descendent pairs.

One approach, suggested by David Lambert, is to find the closest living sequence to each of the ancient sequences and assume that it is a direct descendent. The closest sample meaning the one that differs at the least number of sites in the alignment, identical matches are possible. The majority of ancient samples will have no living descendants within our modern sample, so the ancient sample and its closest modern sample will probably be related through some common ancestor,

rather than a direct line of descent as depicted in figure 3.1. The effective time during which nucleotide substitutions can accrue between a pair of samples is twice the time back to their most recent common ancestor (MRCA), minus the ages of both samples. This means that the effective times between each pair of modern and ancient sequences is likely to be much larger than the age of the ancient sequence as measured by carbon dating. Hence, the method of counting the number of differences between an ancient sequence and the modern sequence closest to it, and then dividing this number by the age of the ancient sample, will tend to overestimate the rate.

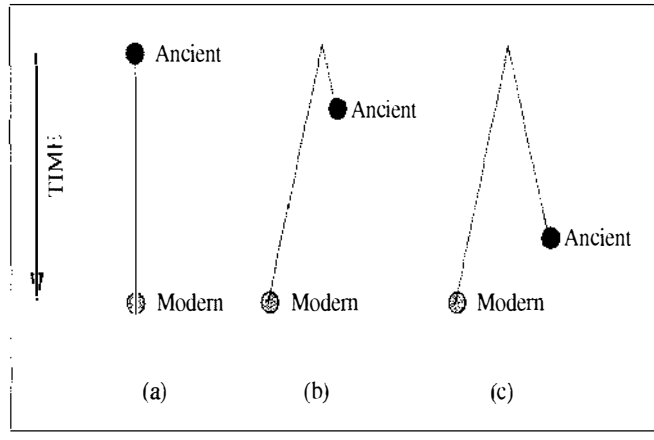


Figure 3.1: The ideal case (a) where the ancient sample has a direct descendant amongst the modern samples is unlikely to occur. It is possible that the effective time between the two samples may be greater (b), or even much greater (c), than the difference in their ages.

Another factor that is not accounted for by the method described above is multiple substitutions at a single site. The number of differences observed between any two sequences is a lower bound on the number of substitutions that actually occurred. This problem is exacerbated by the strong transition bias, meaning that sites may switch back and forward within either the purines {A,G} or the pyrimidines {C,T}. Also, the sites may evolve at different rates [1]. The effect

[1]This begs the question that if there are highly different substitution rates along the DNA molecule, which one are we trying to measure? What we mean to do is measure the **average** rate

of ignoring multiple changes is to underestimate the number of substitutions and hence underestimate the rate.

In order to better understand how biased the direct descent assumption is I conducted a simulation study of the birth/death process within an Adélie penguin population.

## 3.3 Simulations

The aim of this experiment was to use a simulated model of a penguin population over a long period of time to build an empirical distribution of the effective times between pairs of penguins. If a pair of penguins is contemporary then the effective time between them is simply twice the time to their MRCA.

The population is simulated forwards through time from an initial seed population. A sample is then drawn from the modern (simulated) penguins, and the genealogy of this sample is reconstructed using information stored during the simulation. Repeating this process we can derive information on the mean and variation of the effective time between random pairs of penguins. Furthermore, given a model of sequence evolution, we could evolve sequences along the recovered trees. Then the procedure suggested by Lambert, where the closest modern sequence to an ancestral sample is assumed to be a direct descendant, could be tested to estimate by how much it underestimates the effective time between pairs.

A theoretical approach to the same question would be to use the coalescent [55, 56] to derive expected times between pairs. The coalescent is a continuous-time Markov chain that describes the relationships between individuals within a haploid population. The assumptions of the coalescent about population structure are quite strong. In particular, following the description in [56], the assumptions are:

- Discrete non-overlapping generations.

of evolution across all sites of the aligned segment of the HVRI molecule.

- Each individual has exactly one parent in the previous generation.
- Constant population size $N$.
- The number of offspring born to an individual is a random variable subject to the restriction that the sum of offspring born to all the individuals in each generation is $N$.
- The number of offspring per individual follows a symmetric multinomial distribution (under the neutral Wright-Fisher model).

These assumptions are mathematically equivalent to members of a given generation choosing their parents at random from the previous generation [56]. In the context of the coalescent, each generation is equivalent to a breeding season. The first assumption, of non-overlapping generations, is clearly violated in Adélie populations, individual penguins may live and breed over many seasons. In practice, Adélie penguin clutch sizes have never been observed to be more than two, so the last assumption is biologically unrealistic because it does not restrict parents to two chicks per season. It is possible to treat arbitrary distributions of family sizes given an appropriate adjustment of the real population size to an "effective population size" (Ellen Baake, personal communication). However, such adjustments to the standard coalescent are not used here. The second assumption is true given that mtDNA is only inherited maternally, and the third assumption may be a reasonable approximation to the truth over the time-scale of the data (David Lambert, personal communication).

For a constant population of size $N$ the expected time to the MRCA for a random pair of penguins is predicted by coalescent theory to be $N$, with a variance of $N(N-1)$ [56]. We shall see how this expectation compares with the results from our simulations under more realistic assumptions about population structure.

Simulating a penguin population requires information about penguin life-cycles and breeding habits. In particular, approximations are needed for the birth and

death distributions, that is, the probability $p = B(x, a)$ of a penguin laying $x$ female eggs given its age $a$, and the probability $p = D(a)$ of a penguin dying given its age. I used information from two studies on the Adélie life-cycle [63, 109], these studies gave empirical reports of the birth and death distributions, and also some information on colony size which is highly variable from colony to colony (observations ranged from 20–200,000 pairs with many in the range 20,000–30,000) [109].

The following aspects are modelled in the birth/death process.

- Clutch sizes are observed to have one or two eggs in 20% and 80% of cases respectively.
- The sex ratio at hatching is 1:1.
- Age of first breeding has the following distribution: 25% at 4yrs, 25% at 5yrs, 25% at 6yrs and 25% at 7 years.
- 60% of eggs successfully hatch.
- 75% of hatched chicks survive to fledging.
- Adult birds (older than 4 years) have a 10% mortality rate per year.

Table 3.1 is a condensed representation of the points above. Note that because we are only interested in the genealogy of the mitochondrial DNA, and this passes exclusively through the maternal line we need only record the births of female chicks.

The first model was very straightforward, in each year every living penguin has a probability of having zero, one, or two female chicks survive to fledging, and a probability of dying, as summarised in table 3.1. The simulation moved in time steps of one year. This model led to unstable populations that grew very quickly. Penguin colony sizes are observed to be roughly stable, so I altered the model to include a phase that killed off "excess" penguins, and thus kept a limit on the

| Age of Mother | Pr($x$ female chicks survive to fledging) $x = 0$ | $x = 1$ | $x = 2$ | Pr(survival of mother) |
|---|---|---|---|---|
| $3^-$ | 0 | 0 | 0 | 0.700 |
| 4 | 0.825 | 0.125 | 0.050 | 0.900 |
| 5 | 0.650 | 0.250 | 0.100 | 0.900 |
| 6 | 0.475 | 0.375 | 0.150 | 0.900 |
| $7^+$ | 0.300 | 0.500 | 0.200 | 0.900 |

Table 3.1: Birth and Death probabilities used in the simulation of Adélie penguin populations.

maximum number of living penguins. This is possibly a realistic model of predation and limited food resources acting to keep the penguin population in balance.

As the whole population, both living and dead, needed to be recorded I wanted the simulation to be as efficient as possible in terms of space. The penguin population is represented by an array, with a column of three integers for each individual female. For each individual I record: the year each penguin was born, which column contains the penguin's mother, and a third integer that implements a linked list of all the living penguins. On the computers available the upper limit on array size was approximately $700,000 \times 3$. The size of this data structure meant that the maximum population size that could be simulated on the computers available was approximately 300 penguins over $5,700$ years.

The purpose of the simulation was to estimate the distribution of times back to the MRCA for selected pairs of penguins. To do this each selected pair of penguins must have a single common ancestor within the period of time that has been simulated. Hence, the speed at which the population coalesces puts a limit on the size of the population that can be modelled. For instance, the observed proportion of simulation runs for populations of size 300 that did not coalesce within 5,700 years was $1.8 \times 10^{-4}$. Kingman [56] gives a theoretical upper bound on the probability that a population of size $N$ will not have coalesced within $g$ generations of $3(1 - N^{-1})^g$. For $N = 300$ and $g = 5,700$ this upper bound is $1.628 \times 10^{-8}$. So, under more realistic assumptions, the time to a common ancestor

| | Simulation | | Theoretical | |
|---|---|---|---|---|
| Popn. Size | Mean time to MRCA | Standard Deviation | Mean time to MRCA | Standard Deviation |
| 100 | 471.35 | 461.47 | 100.00 | 99.50 |
| 200 | 926.52 | 917.06 | 200.00 | 199.50 |
| 300 | 1428.10* | 1197.26* | 300.00 | 299.50 |

Table 3.2: Means and standard deviations of the time to the MRCA for pairs of contemporary penguins. For comparison both the simulation results, and the values predicted by coalescent theory [56] are shown. (Note that for populations of size 300 not all the samples coalesced, the results were obtained by setting the values for the 0.018% of pairs that had no common ancestor to the maximum possible value of 5,700 years.)

is greater than that predicted by coalescent theory.

I performed simulations with population size limits of 100, 200 and 300. For each generated genealogy I took a sample of 20 living birds and calculated the time to the MRCA for each pair of birds in the sample. This process was repeated 1,000 times for each of the three population sizes. The results are shown in table 3.2.

Given these three simulations ($N = 100, 200, 300$) the relationship between the time to a MRCA for a pair of penguins and population size appears to be linear, as with the theoretical coalescent of Kingman. In each case the mean time in years to the MRCA for pairs of penguins is approximately 4.5 times the population size. However, it may be misleading to extrapolate to realistic population sizes.

Originally I had hoped to extend this simulation to modelling the process of sequence evolution as well. It became obvious that this would greatly exceed the memory resources in the computers I had access to. The simulation is limited to a maximum population size of 300. The field study sampled over 400 penguins from a population that may be as big as 10,000,000. (It is unclear what effect migration between different colonies will have on the effective population size.) This means there is no chance of simulating the sampling process on a realistic scale. Another difficulty with simulating sequences is the large number of parameters that have to be assumed or estimated from the data, for example, the rate, the

transition/transversion ratio, and the distribution of rates across sites. For these reasons the goal of simulating the sequence evolution process was not extended.

Although the population simulation was not totally successful in its aims, it is clear that the behaviour of the simulated population is significantly different from that predicted by the standard coalescent. The distribution of times between pairs of penguins is about 4.5 times larger than the value predicted under the Wright-Fisher neutral model. It would also be interesting to do a comparison to the coalescent where the "effective population size" has been corrected to take into account the observed distribution of family sizes.

## 3.4 Haplotype sampling

The proportion of the existing haplotypic diversity within the Adélie population that has been sampled, is an important quantity to estimate. If most haplotypes have been sampled, then when an ancient haplotype is identified which does not occur in the modern samples, we can infer that this represents a real change in the population over time, rather than a sampling artifact. In this section curve-fitting is used to estimate the proportion of haplotypes that have been sampled.

It is expected that for a given region of DNA, in this case HVRI, that many individuals will share common haplotypes. When the collection process begins many different haplotypes will be found, but as more samples are obtained progressively fewer will be new haplotypes. Eventually, if enough samples are collected, every haplotype in the population will be observed. The discovery curve shown in figure 3.2 represents this process for the Adélie sample collection. The region of sequence that was used is the same trimmed section described in section 3.6.

By fitting a theoretical curve to the one observed in figure 3.2 it is possible to estimate how many haplotypes there are in total, and how many samples would need to be collected before we could expect 90% or 95% coverage of the haplotypes.

A theoretical curve of the form $y = \frac{ax}{b+x}$ (suggested by Allen Rodrigo, personal

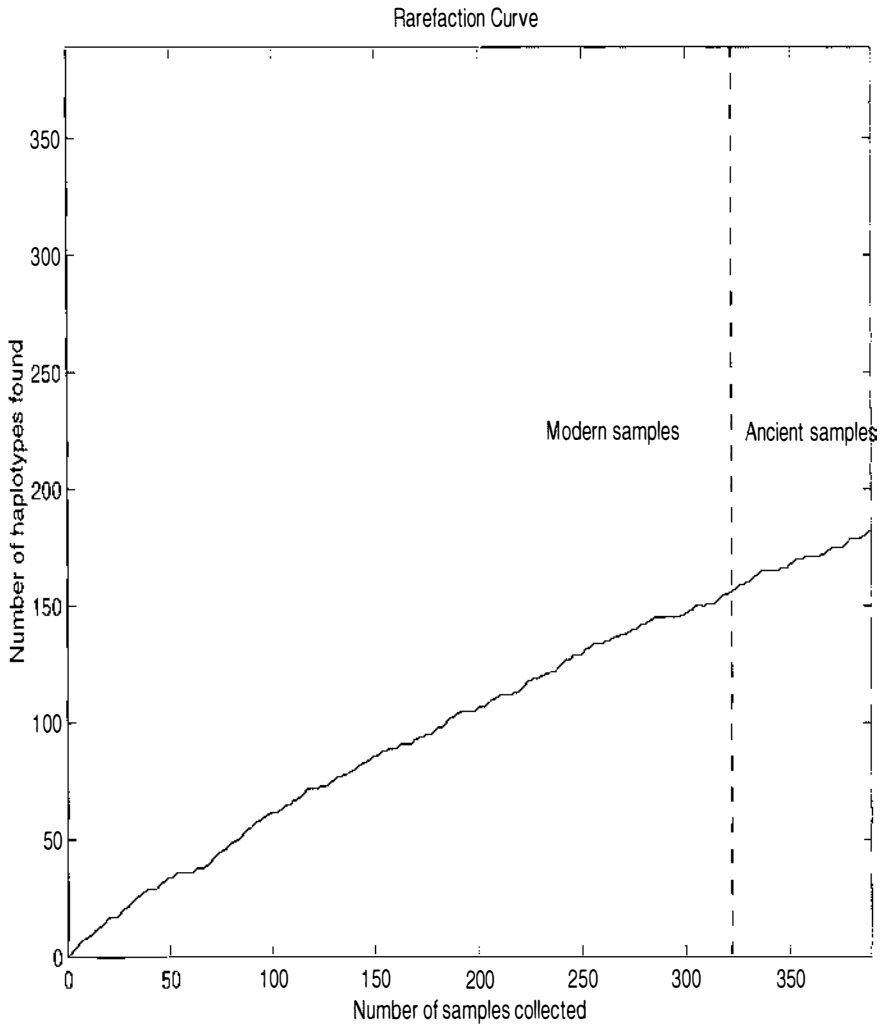


Figure 3.2: The discovery curve for the Adélie penguin samples. The modern samples have been grouped together, followed by the ancient samples.

communication) was fit to the extant samples.

- $x$ is the number of samples that has been collected,

- $y$ is the number of unique haplotypes observed,

- $a$ (the horizontal asymptote of the function) is the estimate of the total number of haplotypes in the population.

In order to fit a smooth curve I randomly permuted the 322 modern samples into 50 arbitrary orders. Each permutation $j \in [1, 2, ..., 50]$ of the samples results in a different vector $\mathbf{y}_j$, where $y_j(i)$ is the number of unique haplotypes observed after $i$ samples have been included, for the $j$th permutation, with $i = 1 : 322$. The component-wise average $\mathbf{y}^*$ of the 50 $y_j$ vectors was calculated,

$$y(i)^* = \sum_{j=1:50} y_j(i)/50, \quad i = 1 : 322.$$

The best fit least-squares curve for $\mathbf{x}$ =1:322 and $\mathbf{y}^*$ was found using numerical optimisation, the curve is shown in figure 3.3. The estimate of the total number of haplotypes is $a = 503.98$. This predicts that we have currently seen only 36% of the haplotypes, and that to see 90%, 6277 samples would have to be collected, or to see 95%, 13252 samples. It should be noted that this is wildly impractical within the current budget of this project, so any hope of getting a near to complete look at haplotype diversity should probably be abandoned,

To estimate a confidence interval for $a$, I used the bootstrap technique [25]. The original set of samples was resampled 100 times with replacement to get a bootstrap estimate of $a$. The mean value of $a$ for bootstrap samples was 204.83 with standard deviation of 29.09. The difference between the bootstrap mean and the observed mean is our estimate of the bias in the bootstrap mean. This value $\widehat{\text{bias}} = 503.98 - 204.83 = 299.17$ was added to each of the bootstrap sample results. The bottom and top 5% of values were discarded to give a 95% confidence interval

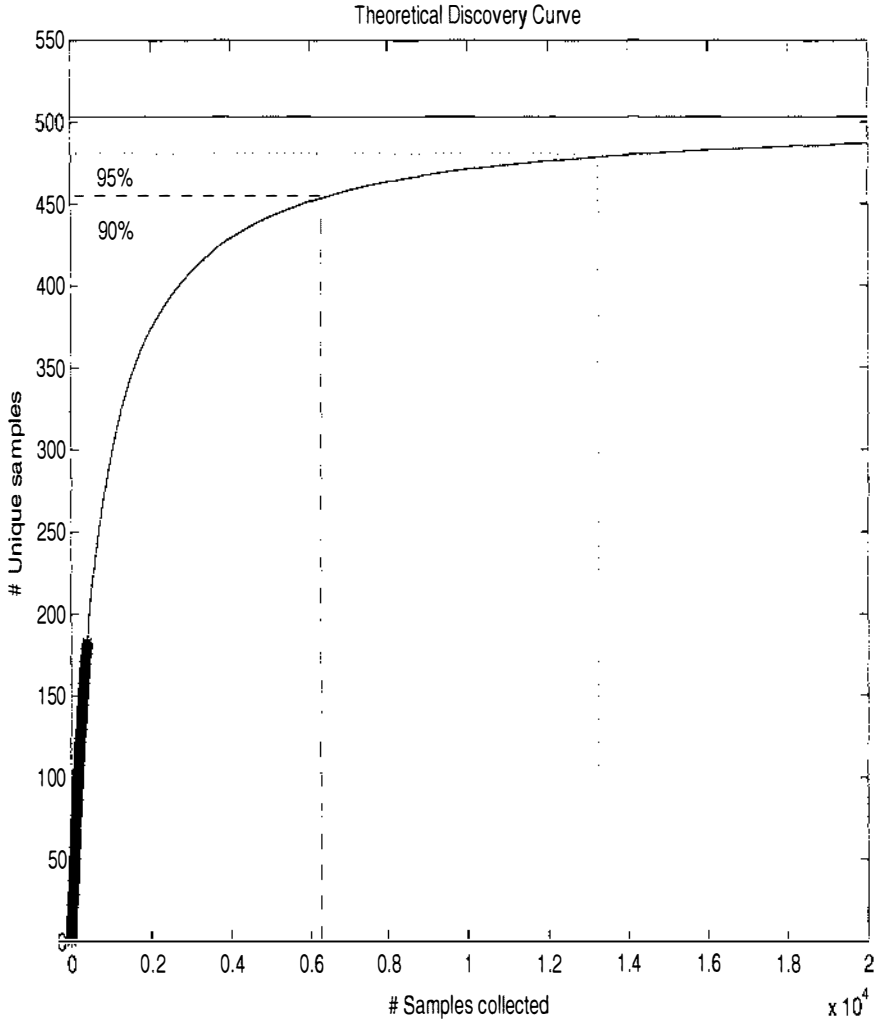


Figure 3.3: The best fit theoretical discovery curve for the Adélie penguins. The observed data is shown by the bold line.

for $a$ of (469.23, 554.99). Note that for the bootstrap to be an effective tool you hope that the bias is close to zero. Here the bootstrap mean is only 40.64% of the observed value for $a$, so the confidence interval should be treated very cautiously, as the standard error has almost certainly been underestimated.

Note also that the choice of theoretical curve is an arbitrary one. Mike Steel (personal communication) suggested an alternative curve of the form $y = a(1-(1-\frac{1}{a})^x)$ which would lead to a different estimate of the total number of haplotypes $a$. The theoretical justification for this curve is that if one places $x$ objects randomly into $a$ boxes the expected number of boxes that have at least one object in them is $y$.

## 3.5 Phylogenetic Analysis of the Adélie Data

The Adélie penguin data set has some general features that are typical of mitochondrial DNA. There is a strong bias towards transitions in the data rather than transversions, the former occurring 63.37 times more frequently. This means that substitutions happen frequently within the purines A and G, and within the pyrimidines C and T, but much less frequently between these groups. The base content was checked to see if there are any trends over time, there is no significant difference in base composition between the ancient and modern populations. Base frequencies estimated from the data are $f_A = 0.3058, f_C = 0.1877, f_G = 0.1970, f_T = 0.3095$.

The control region is expected to have variable rates across sites [65]. A non-consensus plot was constructed to visualise the pattern of variability across the alignment (see figure 3.4), it shows the proportion of taxa at each site that had a different state from that of the consensus sequence. The tall peaks shaded red in the non-consensus plot corresponded to sites that split the taxa into major subgroups, disregarding a small amount of noise, these thirteen sites are mutually compatible and define the tree shown in figure 3.7. Eight of these sites correspond to the same split of the taxa set. This division of the data set is its most pronounced feature.

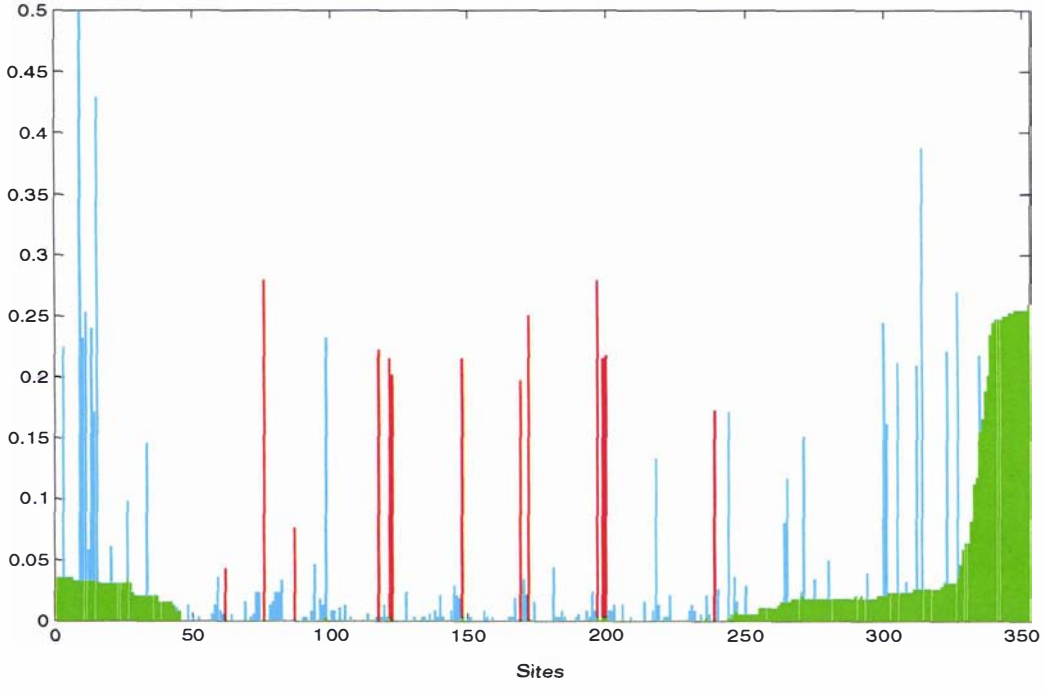


Figure 3.4: The data contains 353 sites over the character set (A,C,T,G,?,-) where a question mark indicates an uncertainty and '-' represents an indel. The bars in this plot show the proportion of non-consensus sites, that is the number of taxa that differ from the consensus sequence at that site, divided by the total number of taxa. The thirteen bars coloured red are consistent with the tree shown in figure 3.7. The green bars represent the proportion of uncertainties at each site.

The same split can be clearly seen in the majority rules consensus tree (figure 3.5). Lambert and Ritchie found that this split was correlated with the geographical location of the penguins. The smaller of the two groups (bottom cluster of figure 3.5) is found only around the Ross Sea, whereas the other group is found all over Antarctica.

The number of uncertainties (missing values) at each site is marked in the non-consensus plot (figure 3.4) with a green bar; these occur more frequently at either end of the alignment, the section between sites 45 and 242 is of very high quality. Within this segment, 123 of the 198 sites are constant, 68 sites display only transitions and 7 sites contain transversions.

It does not make sense to represent the Adélie data by a fully-resolved (binary) tree. With nearly 400 taxa and only 75 informative sites, 68 of which are binary, it would be impossible to infer a fully-resolved tree even if all the sites were compatible. The bootstrap-consensus tree shown in figure 3.5 is not well resolved. If the aim of tree construction is to to identify putative ancestry, then this tree is not a useful representation. The only split that occurs with high frequency is that between the Ross Sea (RS) and Antarctic (A) lineages, most of the other information in the data is lost.

Because the time scales are short, there are commonly a small number of changes between the haplotypes in our sample. The average Hamming distance[2] between samples is 0.0387. The average distance within the A lineage is 0.0182, and within the RS lineage 0.0238. There are several common haplotypes with a number of less frequent haplotypes only one point mutation distant from them. This aspect of the data cannot be represented as a tip-labelled tree. We need flexibility to represent taxa by internal, as well as external, nodes.

Another common feature of the data are pairs of incompatible splits such as displayed by the small data set in figure 3.6 (a). There are two equally parsimonious trees for these sequences, they are shown in figure 3.6 (d). The pattern in (a) must

[2]Calculated over the six letter alphabet {A,C,G,T,-,?}

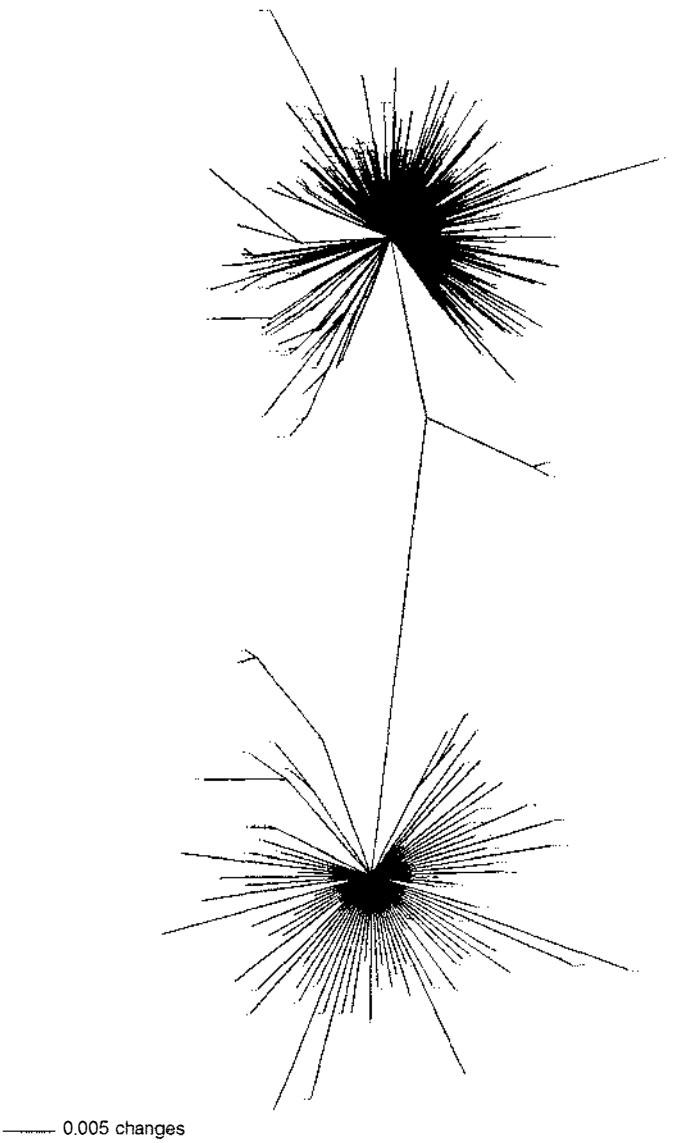


Figure 3.5: The majority-rule consensus tree for the Adélie data. The tree was made using the bootstrap option in PAUP* [97] with 100 replications and uncorrected distances. The tree displays only those edges with a bootstrap support of greater than $\frac{50}{100}$. The lack of resolution reflects the fact that, apart from the major division between the Antarctic (top) and Ross Sea (bottom) lineages, there are very few splits with support at more than one site.

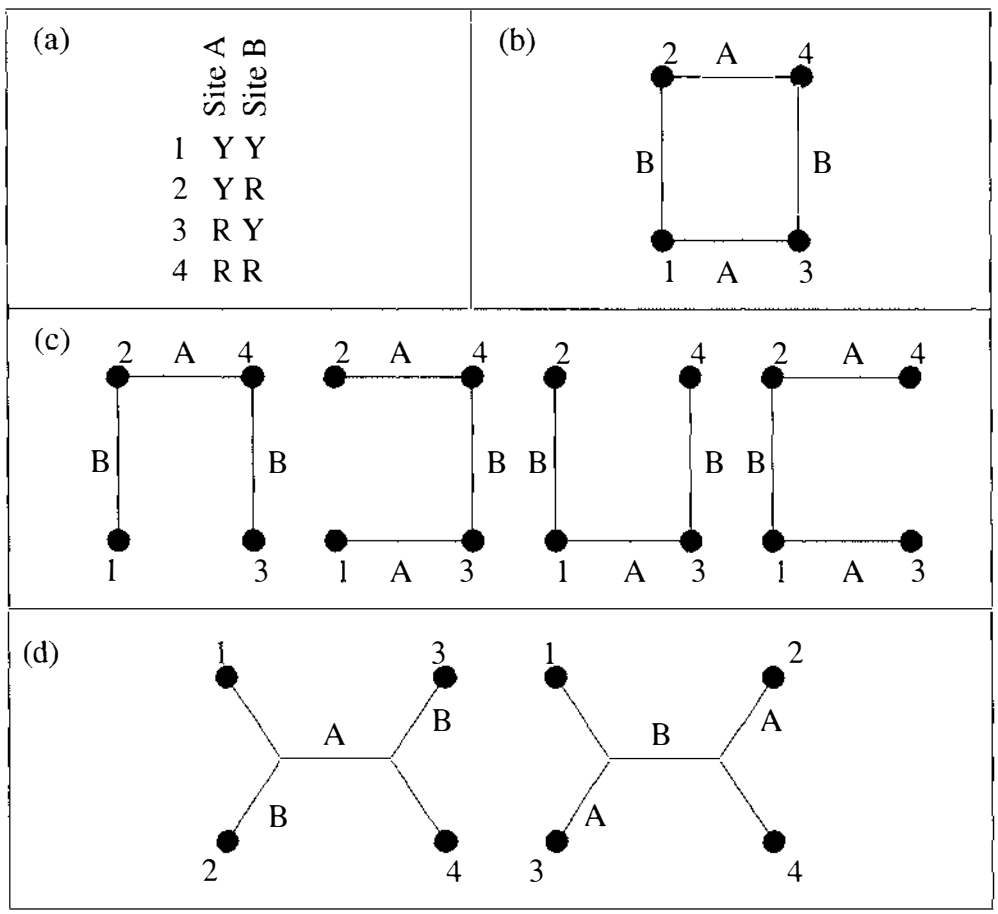


Figure 3.6: Panel (a) shows a common pattern within HVRI. Such a pattern could result from four convergent mutations (b), or three parallel mutations, four possibilities are shown in (c). Panel (d) shows the two most parsimonious trees for the data. Edges are labelled with the site at which a change has occur-ed.

be the result of at least one parallel mutation, but without further information it is impossible to say if the duplicated mutation occurred at the first or second site. Panel (c) shows the four ways of representing such a parallel mutation as a tree where internal nodes are allowed to be labelled. Such a pattern might also be the result of four convergent mutations, as shown in panel (b).

Median networks provide an alternative representation of the data that retains the ambiguity highlighted above. They allow multifurcating graphs where both the tips and internal nodes may be labelled. A pattern like the one in figure 3.6 (a) would result in a cycle within the network that reflects the uncertainty about what

sort of parallel or convergent mutation has occurred. In the next section a median network is constructed for the Adélie data.

## 3.6 Median Networks of the Adélie Data

A median network [3] is a graph that contains labelled nodes and edges. Each node represents a distinct sequence, some of these nodes correspond to observed haplotypes and some are inferred intermediates. If the data contains identical sequences then the node representing these multiple sequences is usually drawn proportionately larger. An edge between two nodes indicates a single difference between these two sequences at a particular site. Parallel edges represent differences at the same site. Formal definitions of median networks, and how they are constructed, can be found in the literature, see for example [3, 8].

Median networks have previously been shown to provide useful representations of intra-species mtDNA data [7, 106]. A median network could potentially be a more useful way of visualising the Adélie penguin data than a binary tip-labelled tree because they:

- Allow labels at internal nodes.
- Can be multifurcating rather than binary.
- Retain uncertainty about parallel or convergent mutations, rather than forcing arbitrary decisions.

A median network combining the modern and ancient samples may give an indication of how much of the observed haplotypic diversity was present in the ancestral Adélie penguin population, and how much has accumulated more recently. It would also be a guide to identifying haplotypes present in the ancient population that are now extinct.

Median networks are designed to work with two-state data. This is not a great limitation with the Adélie data, as the vast majority of the observed changes are

transitions, so most columns in the alignment contain only one or two states (191 of the 198 of sites between site 45–242). Also, data used to build a median network is usually restricted to sites without gaps or missing values.

As most sites in the sequence alignment contained either gaps or missing values, the 15 sequences that contained most of these were removed from the analysis. These sequences were from twelve bones (PE81.Ross.4185, PE93.Pri.old, PE55.Inex.X, PE110.Ade.8030, PE106a.Inex.X, PE106b.Inex.X, PE95.prior.U, PE91.Prior.U, PE75.Ross.4185, PE72.Inex.X, PE69.Inex.X, PE32.Bird.330), and three outgroup sequences (Gento, Chinstrap(2)). Over the remaining sequences the central region of sites 45–242 was virtually free of missing values. Within this region the four sites with gaps and the single remaining site with a missing value were removed. This left a reduced sequence alignment of 193 sites for 67 ancient and 322 modern samples. Of these 193 sites, 119 were constant, leaving 74 discriminating sites, 12 of which were singleton. Over the 62 non-singleton sites there were 152 unique sequences (haplotypes).

This reduced data set is still too large to allow visualisation of the median network. The size of the largest pairwise incompatible set corresponds to the number of dimensions required to fully draw the network. In the data there are groups of at least six mutually incompatible sites, this means that the full median network requires representation as a six, or higher, dimensional hypercube. While this may be an interesting mathematical object, it is hard to represent visually. To simplify, I chose major splits in the data that divided the samples into smaller groups.

The split between the Antarctic (A) and Ross Sea (RS) haplotypes was the most obvious division as it is supported by 8 different sites in the data. This split corresponds to the two main clusters shown in figure 3.5. A sequence was assigned to group RS if it had changes at seven or more of the following eight sites: 117, 121, 122, 148, 172, 197, 199, 200. Within the A group three further sites (62, 76, 169) divided A into four subgroups. In the RS group two further sites (87 and 239), divided this group into three subgroups. These five sites represent five different

splits, and, disregarding a small amount of noise, are compatible with each other and the main A vs RS split.

Together these thirteen sites define the tree shown in the overview diagram (figure 3.7), they are also highlighted in the non-consensus plot (figure 3.4). The criteria for choosing these splits was that: they are mutually compatible (tree-like), and they split the data into reasonable sized groups. Mutually compatible splits that separated only two or three sequences from the rest were not chosen, as they lead to trivially small groups. Of the 389 taxa there were four that did not fit clearly into any of the seven subgroups, removing these taxa meant that the splits were mutually compatible.

The assumptions behind these choices are that these sites are not highly variable, and therefore probably reflect the mitochondrial genealogy of this set of Adélie penguins. Any changes which occur at the same site but in different groups, are assumed to be parallel changes that do not reflect the genealogy (phylogeny). While there is strong support (8 sites) that this is the case for the main A vs RS split, I am less certain about the validity of the assumption for the other splits. Although these splits were each compatible with the main A vs RS split, they were each only supported by a single site.

Median networks were constructed for each of the 7 subgroups: A.1, A.2, A.3, A.4, RS.1, RS.2 and RS.3; as defined above, and shown in figure 3.8. Group A.1 contained a large number of incompatible sites so I have drawn a Minimum Spanning Network [6] for this group, rather than a median network. The reduced data set contains 62 informative sites, six of which have three or more states. Within each subgroup if a site still had more than two states it was excluded from the analysis of that subgroup. The typical pattern seen in the subgroups is to have a central node representing many identical sequences with a radiation of less common types around the central node. We can see that many of the ancient haplotypes are still present in the modern population. The networks display the homoplasy in the data, showing the many mutational pathways that may have occurred.

Further support of the division of the data into the seven subgroups can be seen in the completed median networks (figure 3.8). The central haplotype in each of the subgroups tends to have a much larger number of sequences than those at the tips. Furthermore, these central haplotypes are present in the ancient sample. This suggests a historical division along the lines suggested in the overview diagram (figure 3.7) followed by further radiation around these central types. One interesting observation is that the Gento sequence (an outgroup) has changes at 4 out of the 8 sites used to distinguish the A and RS groups, this suggests that the root of the overview tree shown in figure 3.7 lies on the middle of this long edge (the other two outgroups were uninformative as they contained a long gap through most informative sites). Mid-point rooting also leads to the conclusion that the outgroup is on the long edge between A and RS.

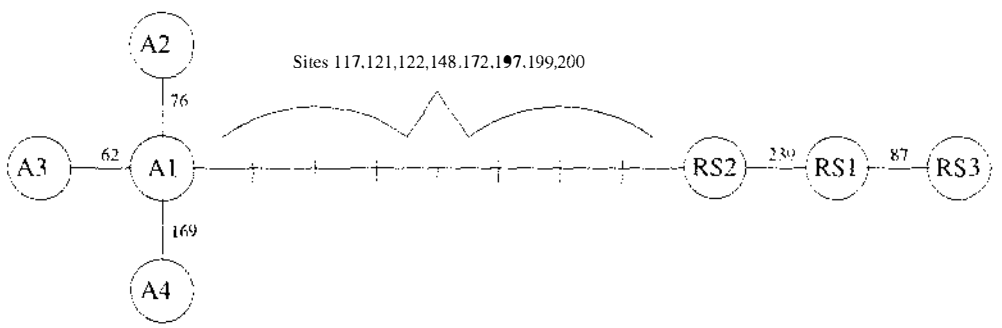


Figure 3.7: The split between the groups A and RS is supported by changes at sites: 117, 121, 122, 148, 172, 197, 199, 200. Within group A there were three mutually compatible splits at sites 62, 76 and 169. These split group A into 4 subgroups; A.1, A.2, A.3 and A.4 Within group RS there were two compatible splits at sites 87 and 239 that split the group into 3 subgroups; RS.1, RS.2 and RS.3.

### 3.6.1 Geographic analysis of subgroups

It had been observed by Lambert and Ritchie [82] that the A and RS groups differed in relative frequency depending on geographical location. I wanted to know if the seven subgroups, A.1, A.2, A.3, A.4, RS.1, RS.2 and RS.3, defined

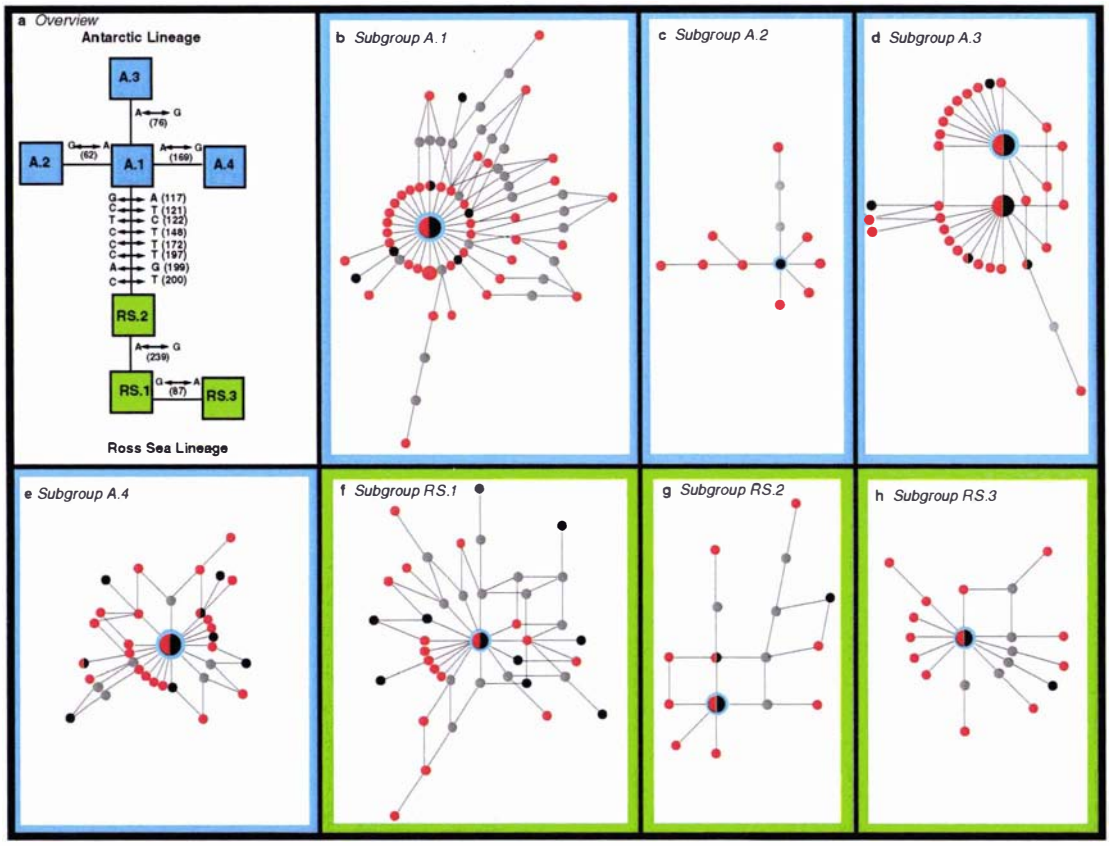


Figure 3.8: The median networks for each subgroup. The red nodes indicate exclusively modern haplotypes, exclusively ancient haplotypes are shown in black, haplotypes found in both modern and ancient sequences are shown in red and black, and inferred intermediates are shown in grey. Small nodes correspond to 1–5 sequences, medium sized nodes 6–20 sequences, and the largest nodes represent $20^{+}$ sequences. In each subgroup there is one node surrounded by a blue circle, this node connects the subgroup to the overview diagram.

| Location | Region | A1 | A2 | A3 | A4 | RS1 | RS2 | RS3 | Total |
|---|---|---|---|---|---|---|---|---|---|
| Inex. Is. | 1 | 6 | 0 | 5 | 7 | 0 | 1 | 2 | 21 |
| Adélie Cove | 1 | 2 | 0 | 5 | 2 | 1 | 0 | 2 | 12 |
| Franklin Is. | 1 | 5 | 0 | 5 | 4 | 0 | 2 | 1 | 17 |
| Cape Bird | 2 | 38 | 4 | 37 | 18 | 10 | 5 | 9 | 123 |
| Beaufort Is. | 2 | 5 | 0 | 9 | 4 | 1 | 1 | 2 | 22 |
| Cape Royds | 2 | 11 | 1 | 15 | 6 | 1 | 3 | 0 | 37 |
| Cape Croz(W) | 2 | 11 | 1 | 3 | 0 | 3 | 0 | 3 | 21 |
| Cape Croz.(E) | 2 | 3 | 1 | 1 | 1 | 2 | 0 | 0 | 8 |
| Cape Hallet | 3 | 5 | 2 | 8 | 2 | 4 | 2 | 4 | 29 |
| Cape W.stone | 3 | 1 | 0 | 1 | 1 | 4 | 0 | 2 | 9 |
| Mawson | 4 | 3 | 0 | 0 | 2 | 0 | 0 | 0 | 5 |
| Davis | 4 | 1 | 1 | 0 | 0 | 0 | 0 | 0 | 2 |
| Ant. P. | 5 | 7 | 0 | 2 | 7 | 0 | 0 | 0 | 16 |
| Total | | 98 | 10 | 91 | 54 | 26 | 14 | 25 | 318 |

Table 3.3: Location of the modern penguin samples by subgroup. Four of the 322 modern sequences do not fall clearly within a subgroup.

before constructing the median networks, also showed any trends with location. If they did it would lend support to the assumption that these subgroups represent the true phylogeny. (Note, a map of the collection sites is shown in figure 1 of Lambert et al. [58].) Table 3.3 records the number of modern samples at each location by subgroup.

The chi-squared test of independence was used to determine if there was any relationship between location and subgroup. For this test to be effective the expected value in each class should not be less than five. This is not the case for this data, however, it is unclear how to best group the different locations, as the classes with low counts tend to be geographically distant from the other locations. I repeated the test with the thirteen locations condensed into five larger regions. To work around the problem of small observed counts I used a randomisation test. The marginal sums for each class were used to define a multinomial distribution, then random samples of the same size as the observed sample were chosen from this multinomial distribution. For each random sample the chi-squared value was calculated and stored. This step was repeated 1000 times. The chi-squared values

| | 13 locations | 5 larger regions |
|---|---|---|
| All subgroups | $\frac{12}{1000}$ | $\frac{8}{1000}$ |
| A vs RS | $\frac{12}{1000}$ | $\frac{0}{1000}$ |
| Subgroups in A | $\frac{73}{1000}$ | $\frac{56}{1000}$ |
| Subgroups in RS | $\frac{188}{1000}$ | $\frac{381}{1000}$ |

Table 3.4: P-values for the test of independence between subgroup and location.

were ranked, and the $p$-values were obtained by counting how many times out of 1000 a larger chi-squared value than the value for the observed data was found in the random samples. Results are shown in table 3.4.

These results agree with the findings of Ritchie [82] that there is a link between geography and the proportions of the A and RS penguins. Within the A group the hypothesis that subgroups are independent of location can be rejected at a level of approximately 5%. However, within RS there does not appear to be any link between location and the RS subgroups.

## 3.7 Calculating the rate

### 3.7.1 Is there a measurable difference in diversity?

It is of general interest to know if the modern samples are more diverse than the ancient samples, disregarding the effect of unequal sample sizes. If they are then it may indicate that the population is expanding, or not yet in equilibrium and that diversity is increasing from some earlier bottleneck.

To address this question, "Is there a difference in diversity?", the seven subgroups were studied individually. Within each subgroup twenty random samples were taken from the modern sequences, each of equal size to the number of ancient sequences in that group. The minimum number of substitutions required to explain the differences between the sequences in each sample, was calculated by finding a minimal length tree, within the network, that connected the sequences. The node

| | Observed Changes Ancient Bones | Mean Changes Blood | STD Blood |
|---|---|---|---|
| A1 | 4 | 9.35 | 3.01 |
| A2 | 0 | 1.50 | 0.22 |
| A3 | 4 | 6.35 | 1.87 |
| A4 | 13 | 11.30 | 1.49 |
| RS1 | 13 | 10.25 | 1.94 |
| RS2 | 4 | 6.55 | 2.89 |
| RS3 | 2 | 2.00 | 1.08 |
| Overall | $40 + 13 = 53$ | $47.30 + 13 = 60.30$ | 5.52 |

Table 3.5: Haplotype diversity of the ancient samples compared to the modern samples. The overall numbers include the 13 changes that separate the different groups. The first column shows the observed number of changes required to connect the ancient samples within each subgroup. The second column shows the mean number of changes required to connect the modern samples over 20 random samples. The third column shows the standard deviation for these 20 samples.

within each subgroup that joins to the overall network, as indicated by the blue circle in figure 3.8, was treated as a root. This root node was always included in each of the trees connecting the random samples.

In only one of the seven subgroups were the modern samples greater than 2 standard deviations more diverse than the ancient samples. The results are summarised in table 3.5. Aggregated over all seven groups the minimum number of inferred substitutions needed to connect the ancient samples (53) was only 1.32 standard deviations away from the mean of the 20 random modern samples (60.3). The conclusion is that there is not evidence to suggest that the ancient samples are significantly less diverse than the modern samples.

### 3.7.2 Using the median network to estimate a rate.

The central question of this project is: "At what rate is Adélie mtDNA evolving?". One way to calculate this rate would be to identify all haplotypes that occur in the modern population, but that were not present in the ancient population at some fixed time in the past. The following method aims to identify such haplotypes using

the median network.

The method that follows is designed for an idealised situation where the modern haplotypes have been completely sampled, so that if a haplotype does not occur in the modern sample we can infer that it does not exist in the modern population. Also, the ancient haplotypes from a specific time point in the past have been completely sampled, so if a haplotype *occurs* in the modern sample but not the ancient sample we can infer that it is the result of a mutation. As was outlined in section 3.4 this is not the case for the Adélie data, it was estimated that only 36% of the modern haplotypes had been sampled. Nevertheless, I apply the method developed below to the Adélie data, and then try to estimate the bias that results from this violation of the underlying assumptions.

### Rate estimation method

The network is a graph $G = (V, E)$, with node set $V$ and edge set $E$. Each ancient sample corresponds to a vertex $v \in V_a$ where $V_a \subset V$. Similarly, each modern sample corresponds to a vertex $v \in V_m$ where $V_m \subset V$. Note that $V_a$ and $V_m$ are not a partition of $V$, their intersection may be non-empty, and there may be elements $v \in V \backslash (V_a \cup V_m)$.

STEP 1 Let $T_a$ be the set of all trees that minimally connect the ancient samples $V_a$.

STEP 2 Let $T_m$ be the set of all trees that minimally connect the modern samples $V_m$.

In other words, find all the most parsimonious trees for the taxa sets $V_a$ and $V_m$. Each tree, $t$ in $T_a$ or $T_m$, will consist of a subset of the edges $E_t \subset E$ in the median network.

STEP 3 Choose a tree $t_a \in T_a$ from the set of trees minimally connecting the ancient samples and a tree $t_m \in T_m$ from the set of trees minimally connecting the modern samples, such that the edges $E_a$ of $t_a$ and the edges $E_m$ of $t_m$ have the maximum possible overlap, i.e., $E_a \cap E_m$ is maximal. (See figure 3.9 for the motivation behind picking trees with the maximum overlap.)

There are four classes of edges in $E$.

(i) Those that are in $E_a$ but not in $E_m$, $E_a \backslash E_m$

(ii) those that are in both, $E_a \cap E_m$

(iii) those that are in $E_m$ but not in $E_a$, $E_m \backslash E_a$, and

(iv) those that are in neither $E_a$ or $E_m$.

The first class of edges, $E_a \backslash E_m$, represent substitutions along lineages that have since gone extinct, they have no direct descendants within the modern population. (Recall that we are assuming the ideal case, where the modern haplotypes have been completely sampled. When applying this method to the Adélie data, where we know the assumption of complete sampling is not true, such lineages may still exist in the population but not have been sampled.) The edges $E_a \cap E_m$ represent diversity that existed in the ancient population and is still present today. Given our assumption of complete sampling, the edges $E_m \backslash E_a$ represent all the substitutions that have occurred along lineages that have not gone extinct, in the time since the ancient population was sampled.

STEP 4 To estimate the rate, divide the number of edges in the set $E_m \backslash E_a$, edges by:

- The age of the ancient samples (we are assuming all existed contemporaneously), in practice these have been collected throughout the last 6000 years rather than from a single period so we substitute the average age of the ancient samples, $\overline{age}$

- The number of sites used in the analysis, $c$, and

- The number of modern samples, $n_m$.

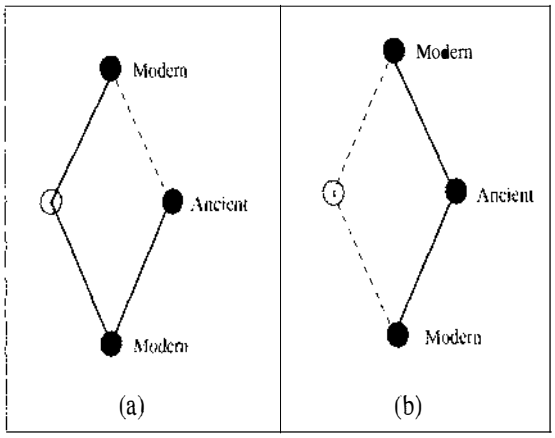


Figure 3.9: In some situations there will be ambiguity as to how much overlap there has been between the substitutions required to account for the ancient samples, and the substitutions required to account for the modern samples. In the situation depicted above we could postulate **either** two changes in the modern samples, plus one change in the ancient samples and no overlap (a) for a total of three changes, **or** two changes in the modern samples, plus one change in the ancient samples and an overlap of one change (b), for a total of two changes. In cases such as this we choose option (b) as it supposes fewer changes overall.

---

$$rate = \frac{\#(E_m \backslash E_a)}{\overline{age} \times c \times n_m}$$

In the Adélie data set it requires:

- 169 changes to account for the changes in the modern samples.
- 51 changes to account for the changes in the ancient samples.
- The maximum overlap is 18 changes.
- So, our estimate of the number of extra changes in the extant sequences is 151.

The average age of the ancient sequences is 2744 years, there are 318 modern samples, and 193 sites were analysed. This gives an estimate of the rate of $151/(2744 \times 318 \times 193) = 0.897 \times 10^{-6}$ or **0.897 s/s/Myr**. Results by subgroup are summarised in table 3.6.

| | $\#E_m$ (Modern) | $\#E_a$ (Ancient) | $\#E_m \backslash E_a$ Edges | #Modern | #Ancient |
|---|---|---|---|---|---|
| A1 | 55 | 8 | 51 | 98 | 13 |
| A2 | 11 | 0 | 11 | 10 | 1 |
| A3 | 29 | 6 | 25 | 91 | 15 |
| A4 | 23 | 14 | 17 | 54 | 20 |
| RS1 | 21 | 16 | 19 | 26 | 10 |
| RS2 | 13 | 5 | 11 | 14 | 6 |
| RS3 | 17 | 2 | 17 | 25 | 3 |
| Overall | 169 | 51 | 151 | 318 | 67 |

Table 3.6: Results of the rate estimation method by subgroup. The set $E_m$ contains the edges in tree $t_m$ which minimally connects the modern samples, the set $E_a$ contains the edges in tree $t_a$ which minimally connects the ancient samples.

This is a point estimate, to get an idea of the variance of this measure I used the delete-one jackknife [25]. For each modern and ancient sequence the number of extra changes required without that sequence in the sample was calculated. Let $\hat{e}_j$ be the number of edges in the set $E_m \backslash E_a$, calculated from the sample excluding sequence $j$. The $\hat{e}_j$ values for the data are shown below.

| Offset from observed $E_m \backslash E_a$ | -4 | -3 | -2 | -1 | 0 | +1 | +2 |
|---|---|---|---|---|---|---|---|
| $\hat{e}_j$ | 147 | 148 | 149 | 150 | **151** | 152 | 153 |
| Frequency | 2 | 2 | 20 | 62 | 289 | 9 | 1 |

$$\hat{e}_j^* = N\hat{e} - (N-1)\hat{e}_j$$

The confidence interval is obtained from

$$\hat{e} \pm t * \frac{s}{\sqrt{N}},$$

where $s = \text{std}(\hat{e}_1^*, \hat{e}_2^*, \ldots, \hat{e}_N^*)$. For the Adélie data set the 95% confidence interval is (0.74, 1.05) s/s/Myrs.

A limitation of this method is that it relies on the assumption that the samples represent all of the haplotypes present in the modern population and all of the

ancient haplotypes that were present at a particular time in the past. Because this assumption is invalid, the method will be biased depending on the respective sample sizes of the modern and ancient populations. If the modern haplotypes have been undersampled, this will tend to reduce the rate estimate, similarly if the ancient haplotypes have been undersampled, then this will tend to increase the rate estimate.

It can be seen from the disparity in the number of changes required to explain the modern samples (169) and the ancient samples (51), that the ancient haplotypes are probably poorly sampled in comparison to the modern samples. Another source of bias is that the method is incapable of returning a negative value, so along with being biased by differing sample sizes there is also an overall positive bias. In order to estimate the size of this bias relative to the estimate, a randomisation test was conducted.

The main obstacle to doing a randomisation test, or any kind of validation, is that it is not possible to write a tractable algorithm for the rate estimation method. As a subproblem the method requires you to find all of the minimal trees on a given set of points within a graph. Finding even a single minimal tree on a subset of the nodes in a graph is known as the Steiner problem within graphs, and is NP-Hard [36]. For this reason it was not possible to do extensive simulations to try and validate the method. However, a small randomisation test was performed for the RS group to try and determine the extent of the bias.

For each of the subgroups RS.1, RS.2 and RS.3, a random sample, of the same size as the number of ancient sequences within that subgroup, was chosen from amongst both the modern and ancient samples. The rate estimation method was carried out treating this random sample as though it contained the ancient sequences, the other sequences were designated modern. This procedure was repeated for 20 random samples in each of the RS.1, RS.2 and RS.3 subgroups.

The distribution of these 20 values for each subgroup, and also the aggregated value over RS, is shown in table 3.7. The mean number of edges in the set $E_m \backslash E_a$,

| | Mean $\#(E_m \backslash E_a)$ | Standard Deviation | Observed $\#(E_m \backslash E_a)$ |
|---|---|---|---|
| RS.1 | 19.85 | 2.54 | 19 |
| RS.2 | 9.10 | 2.00 | 11 |
| RS.3 | 16.45 | 1.28 | 17 |
| Total | 45.4 | 5.55 | 47 |

Table 3.7: Results, by subgroup, of the randomisation test to estimate the bias in the median network rate estimation method.

over all 20 random samples is 45.4, this is our estimate of the bias in the method. The observed number of edges in the set $E_m \backslash E_a$ in RS was 47, so the bias in the method appears to account for 96.6% of our estimate. Correcting the 95% confidence interval for the rate accordingly gives a new interval of (-0.13, 0.19) s/s/Myrs. A negative rate is not possible, so this can be written as [0, 0.19). As the interval includes zero we cannot reject the hypothesis that no measurable evolution has occured over this timescale (6000 years).

Existing approachs for dealing with samples from different periods in time include serial-sample UPGMA (sUPGMA), developed by Drummond and Rodrigo [23], and TipDate [79]. sUPGMA works by first estimating the mutation rate using regression analysis and then adding an extra component to the distances between modern and ancient samples so that the data matrix should become clock-like. A tree is then constructed using UPGMA. TipDate [79] is an extension of the Maximum Likelihood method [33] to cope with non-concurrent samples. The results of these methods on the Adélie data are given in Lambert et. al. [58].

## 3.8 Conclusions

A median network was constructed for the Adélie penguin data. This was a more useful representation of the data than a binary tip-labelled tree, as we were able to model our uncertainty about which substitutions were a reflection of the genealogy

of the sample, and which were a result of homoplasy.

The median network of the ancient and modern samples gave a picture consistent with neutral evolution, in which lineage extinction means that some haplotypes are being lost over time, whilst point mutations cause new haplotypes to be created.

Seven subgroups were identified within the Adélie population that appear to have been present for at least the last 6000 years. The variation within the population occurs as radiations around these central types. The subgroups within the Antarctic lineage appear to be correlated with geographic location.

A method was developed for using the median network to estimate a rate of sequence evolution. The method was discovered to be both biased and computationally intractable, for these reasons I suggest that it is probably not worth pursuing further for the Adélie data, unless a significantly larger number of samples were available. However, it would be interesting to apply the method to a data set where the haplotypes were more comprehensively sampled.

On the data set used in this chapter the sUPGMA and Tipdate methods of rate estimation also give confidence intervals including zero for the Adélie data. Since the time of doing this work a larger number of ancient samples have been collected. Recently, a Monte Carlo Markov chain approach [22] has been applied to this extended data set that gives confidence intervals for the rate that exclude zero [58].

# Chapter 4

# $\delta$-plots: A tool for visualising tree-likeness

## 4.1 Introduction

Within the field of phylogenetic analysis there exist many recipes for turning distance matrices into trees, but there are comparatively few tools available for assessing how appropriate this process may be. As there is no *a priori* reason that a distance matrix should be well represented by a tree, such tools would be useful. In this chapter I present a method, based on statistical geometry [19, 28, 29], that aims to quantify how tree-like a given data set is. The measures of "tree-likeness" for each quartet in the data are combined and displayed visually in a graph we call a $\delta$-plot.

Other methods for assessing the different signals in a phylogenetic data set before a tree has been estimated include RASA (Relative Apparent Synapomorphy Analysis) [61], spectral analysis [44], and likelihood mapping [95]. In a recent paper by Nieselt-Struwe and Haeseler [68] likelihood mapping was extended to a more generalized form called quartet mapping that is also an extension of statistical geometry. Another method is split decomposition, which [5] decomposes an input distance metric into a set of weakly compatible splits, that are not restricted to

forming a tree.

Many processes such as recombination, reassortment, gene conversion, and lateral transfer can lead to reticulate evolution, that is better described by a network than a tree. In this chapter I explore the potential of the $\delta$-plot method to identify reticulate data sets. Furthermore, the method is extended to rank taxa in the data in order of how much they confound the tree-like signal, this may allow the identification of individual recombinant sequences.

The methods are tested on a range of simulated data sets, and three biological data sets: a set of mammal mitochondrial genomes [72], a viral data set [18], and restriction fragment length polymorphism (RFLP) data from *Candida albicans* [86]. These three examples are listed in order of the amount of reticulation that is expected. An extended case study of *Candida albicans* illustrates both the usefulness of $\delta$-plots, and some of their limitations.

The research presented here was initiated during a visit to the Mid Sweden University, Sundsvall, in July 2000, and was supported by a STINT grant. It was developed in collaboration with Vincent Moulton and Katharina Huber. The project was overseen by Vincent Moulton, each of us contributed to the ideas. I had sole responsibility for implementing the simulations and performing the tests on real data.

## 4.2 Background

A common input to phylogenetic problems is a set of taxa with a matrix containing the pairwise distances between them. Example sources for distances include: Hamming distances between pairs of sequences; or, distances inferred from the Hamming distances under some model of sequence evolution; and metrics based on the presence or absence of bands in RFLP data.

Firstly, we define a distance **metric** on the taxa set $X$:

DEFINITION: 2 (METRIC) *Let $X$ be a finite set. A metric $D : X \times X \to \mathbb{R}$ satisfies the conditions:*

1. $d(x, y) \geq 0$, $d(x, x) = 0$, *for all* $x, y$
2. $d(x, y) = d(y, x)$, *for all* $x, y$
3. $d(x, z) + d(z, y) \geq d(x, y)$, *for all* $x, y, z$

Note that we *do not* require that $d(x, y) = 0 \Rightarrow x = y$. In phylogenetic data sets we allow for the case where different taxa are identical on the characters that have been observed.

DEFINITION: 3 *A metric on $X$, is said to be* **additive** *if it fits exactly on some tree. That is, a tree can be constructed, with non-negative edge lengths, and the taxa set mapped into the vertex set, where the sum of the edge lengths along the path between each pair of taxa corresponds to the distance between those taxa.*

Buneman showed that a metric is additive if and only if it satisfies the four-point condition (**FPC**) [12], (see also [4]).

DEFINITION: 4 (FPC) *For all subsets of four taxa $\{u, v, x, y\} \subset X$, of the three pairwise sums: $d(x, y) + d(u, v)$, $d(x, u) + d(y, v)$, and $d(x, v) + d(y, u)$ the largest two are equal. Figure 4.1 illustrates why this is true for additive (tree-like) data.*

Standard methods for constructing trees from distance matrices, such as neighbor-joining [84] and UPGMA [89], do not test the input data for additivity, they will construct a tree from any input distance data, even when the data is random, or generated from a non-tree model. Hence, it is useful to do some prior analysis of the data to see how tree-like it is. This will give an indication of how much confidence should be placed in the resulting phylogeny. If the data is not tree-like under any transformation, there could be other features to the data that are obscuring the historical tree-like signal, such as selection, mutation pressure, or

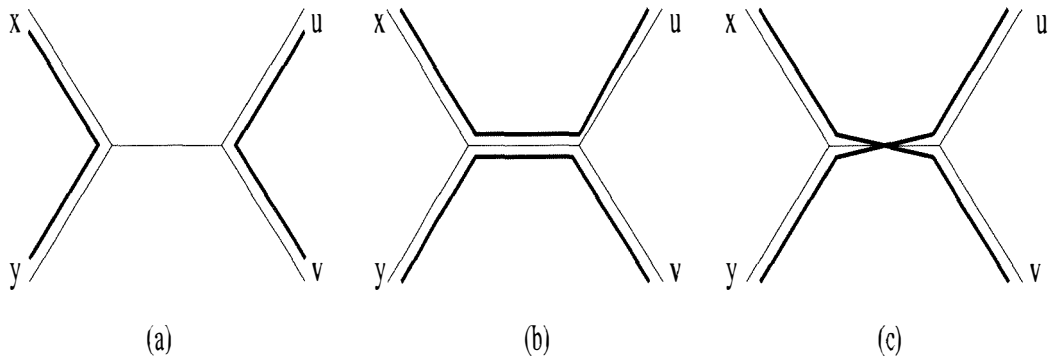


Figure 4.1: There are three possible ways of summing pairs of distances on four taxa: $d(x,y)+d(u,v)$, $d(x,u)+d(y,v)$, and $d(x,v)+d(y,u)$, as shown in (a), (b) and (c) respectively. If the distances are additive then the two largest of these three sums must be equal. In this case the sum of the path lengths in (b) and in (c) are equal and larger than the sum in (a).

systematic error. Alternatively, the data may not have been generated on a tree, but instead on some reticulate network.

Biological data will not normally satisfy the FPC. In the case of Hamming distances, effects such as parallel mutations and multiple changes at a site lead to distances where the FPC is violated. A binary decision where a data set is tested to see if it satisfies the FPC or not is unlikely to be useful except for the most "well-behaved" of data. In the words of David Penny (personal communication), "Unfortunately organisms do not evolve in order to leave a historical signal, their aim is to survive!" In statistical geometry [19, 28, 29] one attempts to measure by how much data departs from being additive. This is done as follows:

Any distance metric on four taxa can be represented by the diagram shown in figure 4.2 for some permutation of $x$, $y$, $u$ and $v$ with an assignment of non-negative weights to the edges $a, b, c, d, s$ and $l$ ($s \leq l$) such that the sum of the edge lengths along a path between each pair of taxa equals the distance between those taxa [4]. If the metric is additive then the shorter internal edges ($s$) will be equal to zero, leaving a tree. If the short edges ($s$) have small weight relative to the long edges ($l$) then the distances for the quartet may be said to be *tree-like*, so the ratio $\frac{s}{l}$ is

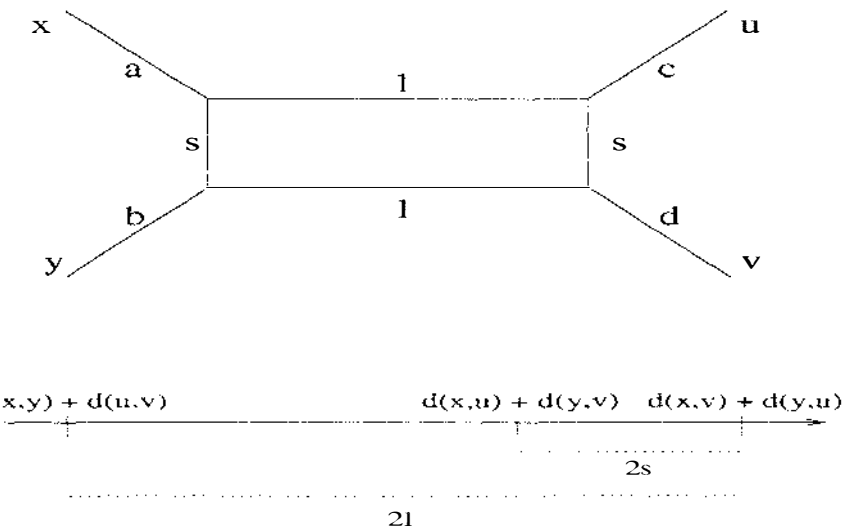


Figure 4.2: Any metric on four points can be fit to the following diagram. The three pairwise sums $d(x,y)+d(u,v)$, $d(x,u)+d(y,v)$, and $d(x,v)+d(y,u)$ are shown in order of magnitude along the real number line. $(d(x,v)+d(y,u))-(d(x,u)+d(y,v)) = ((a+l+s+d)+(b+l+s+c))-((a+l+c)+(b+l+v)) = 2s$. Similarly, $(d(x,v)+d(y,u))-(d(x,y)+d(u,v)) = 2l$. Our measure of tree-likeness $\delta = s/l$.

an indicator of the tree-likeness of the data (see Eigen et al. [29]).

DEFINITION: 5 *We define $\delta$, our measure of tree-likeness for a quartet, to be $\frac{s}{l}$, where $s \leq l$. For every quartet $u,v,x,y \in X$, $\delta(x,y,u,v)$ is defined as $\frac{s}{l}$, where,*

$$\delta(x,y,u,v) = \frac{s}{l} = \frac{\max(\alpha,\beta,\gamma) - \text{mid}(\alpha,\beta,\gamma)}{\max(\alpha,\beta,\gamma) - \min(\alpha,\beta,\gamma)},$$

*$\alpha = d(x,y)+d(u,v)$, $\beta = d(x,u)+d(y,v)$, and $\gamma = d(x,v)+d(y,u)$, and max, mid and min are respectively the largest, middle, and smallest of these three values. (See the caption of figure 4.2 to justify this formula.) Note, in the case where $\alpha = \beta = \gamma$, $\delta$ is defined to be zero, as this corresponds to a star tree.*

$\delta$ ranges between 0 and 1, a value of 0 means the quartet is additive, larger values are progressively less tree-like. The reason it is called *statistical* geometry is that $\delta$ can be calculated for each of the $\binom{n}{4}$ quartets, the mean value of $\delta$ over all quartets, $\bar{\delta}$, is taken as an indication of the tree-likeness of the data as a whole.

For future reference I also introduce here the notation $\overline{\delta}_x$, defined to be the mean value of $\delta$ for all quartets that include the taxon $x$.

In the following section (4.3) I discuss the development of $\delta$-plots, a graphical representation of the $\delta$ values, for a large set of taxa. Section 4.4 deals with simulations testing the dependency of $\delta$ on the number of taxa, sequence length, and the amount of reticulation. In section 4.5 the $\overline{\delta}_x$ values of individual taxa are used to identify recombinants, and I discuss simulations testing the effect of removing the taxa with the highest $\overline{\delta}_x$ values on the ability of neighbor-joining to represent the observed metric on the remaining taxa. The last section of this chapter (4.6) is a case study using the methods developed to explore the theory that the yeast *Candida albicans* consists of sub-species with different modes of reproduction.

## 4.3 $\delta$-plots

$\delta$-plots are a visual tool to evaluate the tree-likeness of a data set, the input required is a distance matrix on the set of taxa, $X$. $\delta(x, y, u, v)$ is calculated for every quartet $u, v, x, y \in X$, and these values are then plotted in a histogram. The number of quartets in a data set with $n$ taxa is $\binom{n}{4}$, so the computational cost of constructing a $\delta$-plot is $O(n^4)$. For large $n$ (say $n > 100$ taxa), it may be satisfactory to construct a $\delta$-plot for a random subsample of the quartets, rather than computing $\delta$ for every quartet.

Figure 4.3 compares the $\delta$-plots for a randomly generated set of symmetric distances constrained to obey the triangle inequality, with a set of distances from sequences generated along a tree. Both data sets contain $n = 30$ taxa.

The random distances were constructed by assigning each entry in an upper triangular distance matrix, a uniformly generated random number between 1 and 2. This ensures that the triangle inequality holds, as the sum of two numbers greater than or equal to one must always be greater than or equal to two, so $d(x, y) + d(y, z) \geq d(x, z)$ for all $x, y, z \in X$.

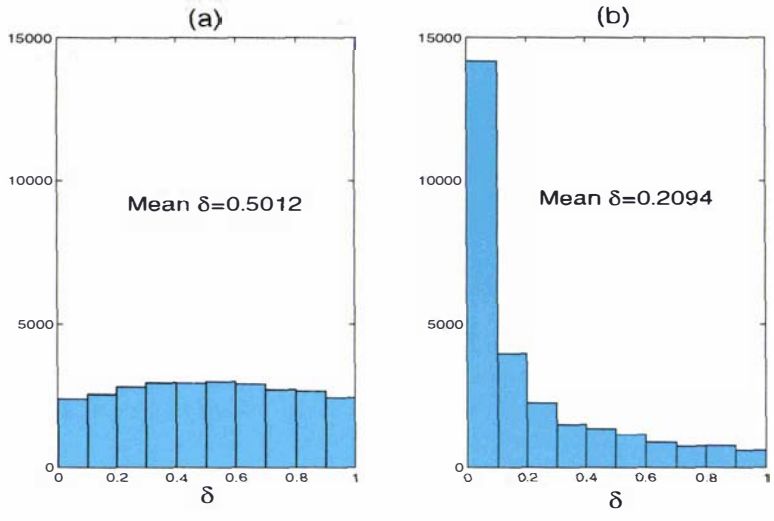


Figure 4.3: Histograms of the $\binom{30}{4}$ $\delta$ values for all quartets, plot (a) is from random distances forced to obey the triangle inequality, and plot (b) is from corrected Hamming distances between sequences generated along a tree. The axes have been scaled to the same length. For random data the expected value of $\overline{\delta}$ is 0.5, [29], if we assume that biological data will behave no worse than random data, then $\overline{\delta}$ should vary between 0 and 0.5.

---

The sequences, of length 1000bp, were generated using *Treevolve* v 1.32 [80], (the sample input file in section 4.4 gives details on the parameters used). These sequences were converted to distances using the *phylip* package *dnadist* [34], with a correction for multiple substitutions according to the model used to generate the data. As the second data set was generated on a tree, we expect that most of the $\delta$ values will be close to zero. However, random noise introduced by the finite sequence length, and the effect of multiple changes mean that we do not expect the distances to be exactly tree-like with $\delta$ values all zero. In figure 4.3 we see that the distribution of the $\delta$ values for the random data (a) is almost flat with a mean of 0.5012, whereas the plot for tree-like data (b) is skewed towards zero with a mean of 0.2094.

In figure 4.4 the $\delta$-plots for three biological data sets are compared. The first data set (a) is taken from complete mitochondrial sequences of 30 mammals [72], (b) is the AIDS virus data set from Debyser et al. [18] containing 21 isolates,

and (c) is yeast data based on RFLP of 266 *Candida albicans* isolates taken from Schmid et al. [86]. The mammal data (a) is based on whole mitochondrial genomes and we expect a strong tree-like historical signal. The virus data (b) may contain recombinant taxa and therefore the tree-like signal may be partially obscured. The yeast data set (c) is based on restriction fragment data which is often found to be a less precise measure of similarity between taxa (see discussion in Swofford et al., pg 412 [96]). Also, it is not known to what extent *C. albicans* reproduces sexually versus clonally; if there is considerable sexual reproduction, then reassortment of the six chromosomes would result in a network rather than a tree signal (although some tree-like signal may remain due to linkage along the chromosomes). For these two reasons we anticipate that the *C. albicans* data may not be very tree-like. These expectations are supported by the $\delta$-plots shown in figure 4.4, with $\overline{\delta}$ values of 0.1611, 0.3017, and 0.4117 for the mammal, virus and yeast data sets respectively.

The $\delta$-plots of both the simulated and the biological data sets exhibit the behaviour that was predicted, that is, the stronger the tree-like signal is, the more the distribution of the $\delta$ values of their quartets are skewed towards zero.

## 4.4 Simulations

Recombination is an important feature in the evolution of some organisms, especially viruses. A number of methods have been developed to identify recombination in general and also to search for breakpoints along recombinant sequences where the "parent" sequence changes. See, for example, [64] for a sliding window approach, phylogenetic profiles [107], and [39] for a likelihood based approach. The method of split decomposition [5], as implemented by Daniel Huson in the program *SplitsTree* [50, 21], can also be useful in analysing potentially recombinant data. It does not force the data onto a tree but instead decomposes the metric into weakly compatible splits plus some residue. If the data is reticulate then this may show up as "boxes" within the *SplitsTree* graph. *SplitsTree* has the limitation that for

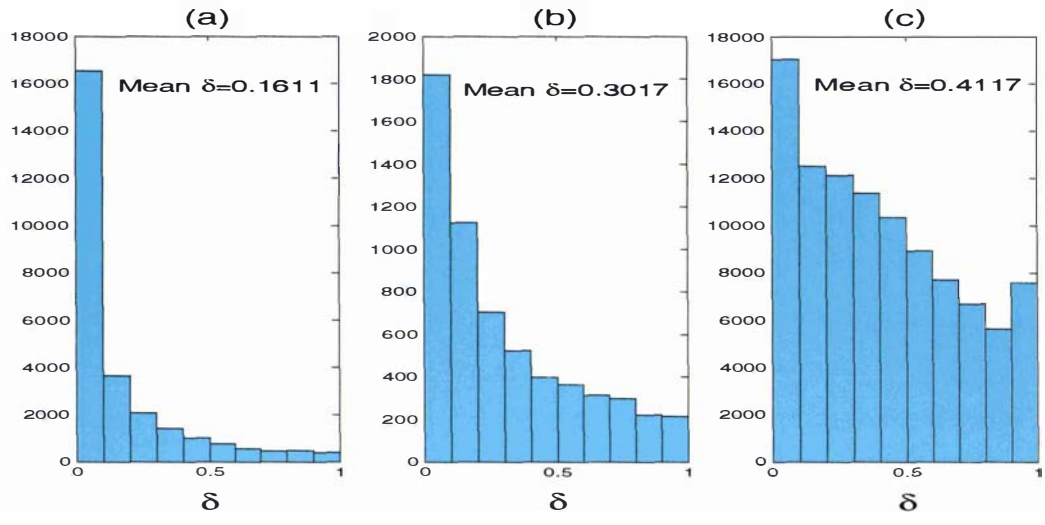


Figure 4.4: The $\delta$-plots for a mammal (a), a viral (b), and a yeast (c) data set. To generate the yeast plot (c) I sampled 100,000 quartets at random from the $\binom{266}{4} = 203,927,570$ possible quartets. Plots (a) and (b) are of the complete set of quartets for these data.

large data sets it tends to produce unresolved star-trees. To investigate if $\delta$-plots may be a useful tool for identifying data sets that result from reticulate evolution, simulations were carried out to test the dependence of $\bar{\delta}$ on the number of taxa, the sequence length, and the frequency of recombination events.

The test data was generated using the programs *Treevolve* version 1.32 [80] and *dnadist* which is part of the *phylip* package. *Treevolve* generates sequences according to a network model where the user specifies the probabilities of both recombinant and coalescent events. Firstly, a network is generated by working from the tips, backwards in time towards the root of the tree. At each time step there is a probability that two taxa will coalesce into a single ancestral taxon; and a probability that a single taxon will split into two, the latter being a recombination event. When a recombination event occurs, a breakpoint is chosen at random along the length of the alignment. When all the sequences have coalesced to a single ancestor, an ancestral sequence is defined at the root, and sequences are evolved forwards in time along the network. Each site in the sequence will have a unique

tree representing its history, depending on which side of the various breakpoints it is. A sample file with the input parameters to the program is shown below. Note that the recombination rate is specified per nucleotide, rather than over the whole sequence.

### A sample input file for Treevolve

```
BEGIN TVBLOCK
        [sequence length] l1000
        [sample size] s30
        [mutation rate] u0.000001
        [number of replicates] n1000
        [substitution model] vHKY t2.0
        [generation time/variance in offspring number]
        b1.0 [1.0 = Wright-Fisher]

        *PERIOD 1
                [population size] n1000000
                [recombination] r0.00000000075
        *END
```

In all the simulations in this chapter the model of sequence evolution used was HKY [41] with a transition-transversion ratio of $\kappa = 2$. Distance matrices were formed using the package *dnadist* which calculates the Hamming distances between sequences and then corrects these according to the specified model (in this case HKY, $\kappa = 2$).

In the first simulation we tested the dependence of $\overline{\delta}$ on $n$, the number of taxa. As $\delta$ is evaluated for each quartet independently, I did not expect that $\overline{\delta}$ would depend on $n$. Sequence length ($c$) was fixed at $c = 500$bps. One hundred repetitions were performed for each value of $n$ in the range $n = (5, 10, ..., 90, 95)$. The number of taxa was found to have no significant effect on $\overline{\delta}$, as is shown in figure 4.5.

In the next simulation (results shown in figure 4.6), the number of taxa was held constant at $n = 30$ and the recombination parameter ($r$) of *Treevolve*, and sequence length ($c$) were varied. I would expect $\overline{\delta}$ to be negatively correlated with

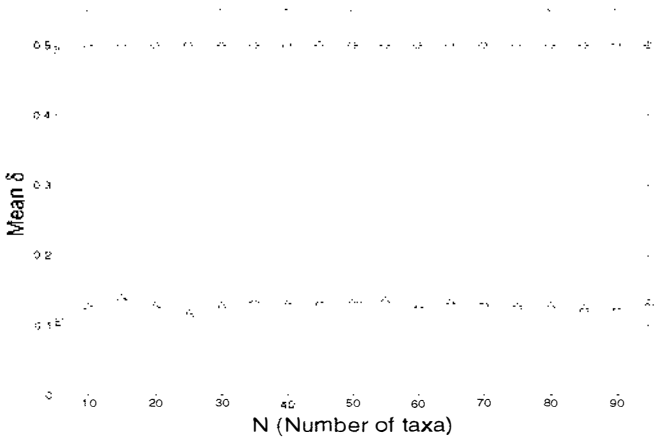


Figure 4.5: $\overline{\delta}$ for increasing $n$, plotted for random(o) and tree-like($\triangle$) data. The top line is from random distances, the lower line is from corrected Hamming distances of sequences simulated along a tree. One hundred repetitions were performed for each number of taxa in the range $n = (5, 10, 15, ..., 95)$. Sequence length was fixed at $c = 500$bps.

sequence length, as in longer sequences, there is a smaller noise to signal ratio. This is certainly the case for zero recombination. For the highest level of recombination shown, $(r = 1 \times 10^{-9})$, $\overline{\delta}$ increases slightly from 0.160 with sequence length $c = 600$ to 0.164 with $c = 800$. An explanation for this is that the recombination frequency parameter, $r$, is specified per nucleotide, so there is more opportunity for recombination to occur in a longer sequence. Overall, the plot shows a positive correlation between $\overline{\delta}$ and the frequency of recombination events.

A similar simulation was done, again with $n = 30$, where $\overline{\delta}$ was recorded for different sequence lengths, and values of $rc$. For a given value of $rc$, the expected number of recombination events is the same for each sequence length. Results are shown in figure 4.7. $\overline{\delta}$ is negatively correlated with sequence length, and positively correlated with $rc$.

The simulations indicate that $\overline{\delta}$ is independent of the number of taxa, so high values of $\overline{\delta}$ suggest either a lack of tree-like signal in the data, or that the signal has been obscured by other processes such as recombination, or biases in the process of sequence evolution. Other factors, that have not yet been tested, such as choosing

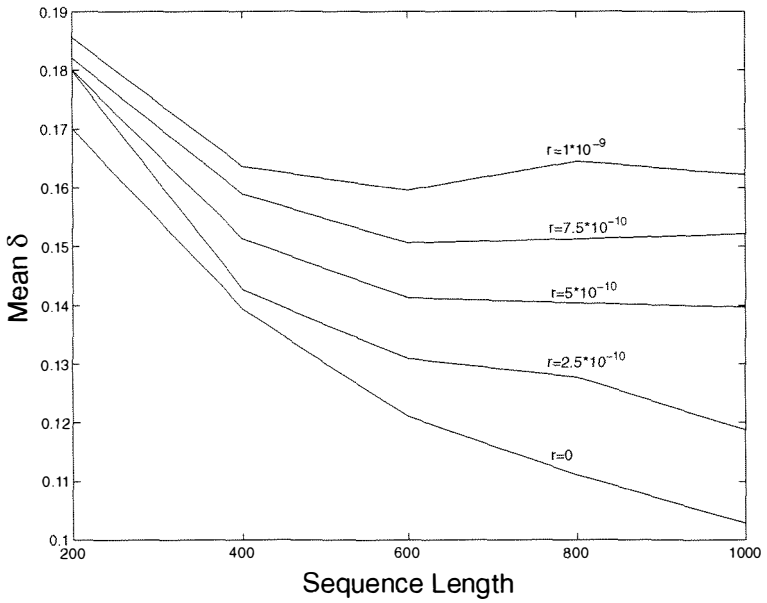


Figure 4.6: $\overline{\delta}$ plotted against sequence length for five different levels of recombination: $r = 0$, $r = 0.25 \times 10^{-9}$, $r = 0.5 \times 10^{-9}$, $r = 0.75 \times 10^{-9}$, and $r = 1 \times 10^{-9}$. The number of taxa $n = 30$ is fixed. Each point is an average over 1000 repetitions.

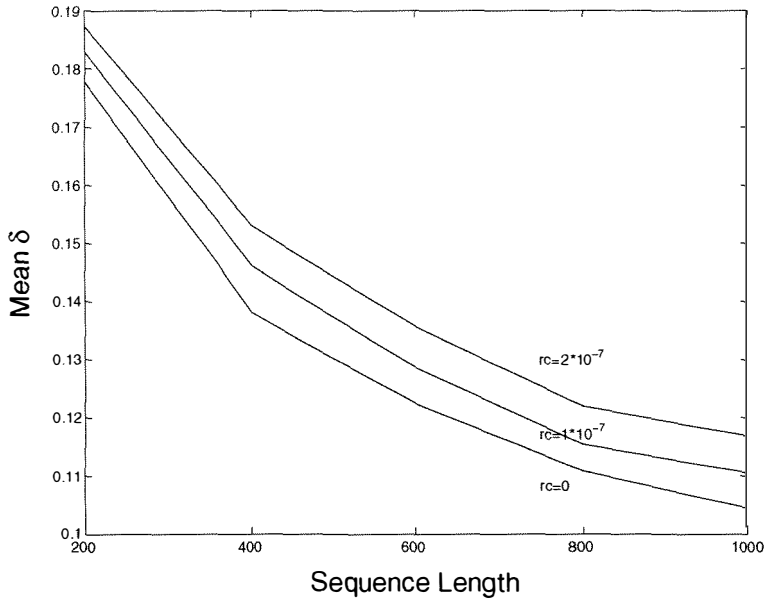


Figure 4.7: $\overline{\delta}$ plotted against sequence length for three different levels of recombination per sequence: $rc = 0$, $rc = 1 \times 10^{-7}$, and $rc = 2 \times 10^{-7}$. The number of taxa $n = 30$ is fixed. Each point is an average over 1000 repetitions.

the wrong model to infer distances by, incorrect alignment, and the underlying tree topology, could also influence $\overline{\delta}$.

## 4.5 Identifying "troublesome" taxa

If the data is tree-like we would not expect any taxon, or group of taxa, to consistently occur in quartets with high $\delta$. However, if a subset of the sequences have (i) undergone reticulation, (ii) been involved in a sequencing or alignment error, or (iii) are at the end of a long edge and have hence become randomised with respect to the other taxa, then we might expect them to appear in a significant proportion of the quartets with high $\delta$. To identify specific taxa that may be obscuring tree-like signal in the data we focus on $\overline{\delta}_x$, the mean value of $\delta$ for all quartets containing the taxon $x$.

All simulations and examples in this section were also performed with a second measure that focused on the tail of the $\delta$ distribution. For each taxon $x$, the number of times that $x$ appears in a quartet with $\delta > 0.95$ was counted. This measure was used in place of $\overline{\delta}_x$. In all cases the results were similar, but in the simulations $\overline{\delta}_x$ was found to be less variable, and a better discriminator between recombinant and non-recombinant taxa.

Example plots of $\overline{\delta}_x$ are given in figure 4.8 for the mammal and virus data sets, previously examined in figure 4.4. The guinea pig sequence from the mammal data set appears to be the only outlier. On excluding the guinea pig sequence, the value of $\overline{\delta}$ was reduced by 7.5% from 0.1611 to 0.1469. It seems unlikely that the guinea pig mitochondrial genome sequence is recombinant, as recombination has never been detected in mammal mitochondria; nor does the guinea pig appear to be on a long branch. A third possibility is that the sequence contains a reading or alignment error. The alignment of the guinea pig sequence should be reexamined, as an error here could explain the high $\delta$ values for this sequence.

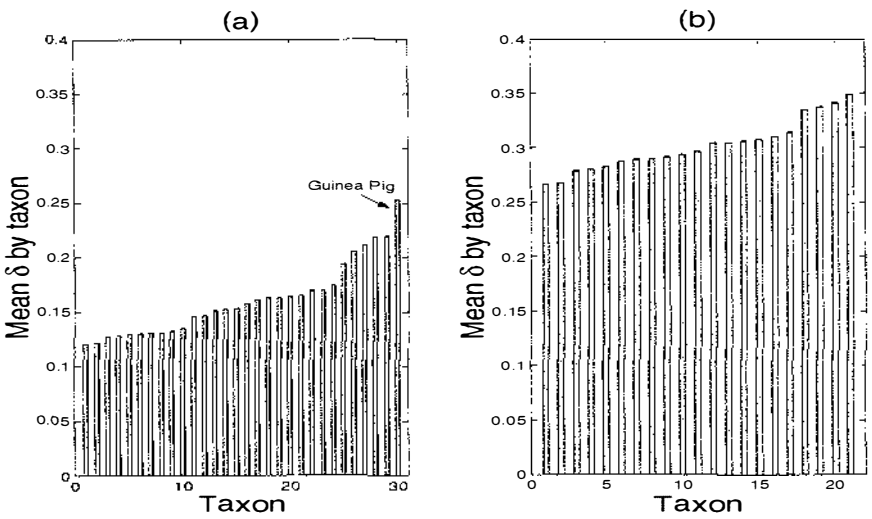


Figure 4.8: $\overline{\delta}_x$ for the mammal (a), and virus (b) data sets. Each column, $i$, represents the mean value of $\delta$ for all quartets that contain the taxon $i$. The columns have been sorted by value. The virus data set has no outliers, but in the mammal data, the guinea pig sequence appears in a slightly larger number of quartets with high $\delta$ values than any other taxon.

### 4.5.1 Removing "troublesome" taxa

Given that we can order the taxa in a data set according to $\overline{\delta}_x$, the mean $\delta$ value of the quartets to which they belong, it is interesting to see how measures of tree-likeness are effected by removing the taxa with the highest $\overline{\delta}_x$ values. The following simulations aimed to determine if removing these taxa significantly improves the tree-likeness of the data, as compared to removing an equivalent number of randomly selected taxa. The following four statistics were used as measures of tree-likeness: $\overline{\delta}$; and the $L_\infty$, $L_1$, and $L_2$ norms between the observed metric $D$, and the metric $D_T$ induced by the neighbor-joining tree, $T$.

DEFINITION: 6 ($L_p$ NORMS)

$$L_p = \left[ \sum_{i,j} |D(i,j) - D_T(i,j)|^p \right]^{1/p}$$

*where $D$ is the pairwise distance matrix on the set of taxa and $D_T$ is the distance*

| | |
|---|---|
| Sequence Length ($c$): | 100 and 500 |
| Number of Taxa ($n$): | 20 and 30 |
| Frequency of recombination ($r$): | 0, $0.5 \times 10^{-10}$ or $1.0 \times 10^{-10}$ |

Table 4.1: Parameters for the removal order simulation. There were $2 \times 2 \times 3 = 12$ experiments. The frequency of recombination is set by the parameter $r$ in *Treevolve* (see the sample *Treevolve* input file in section 4.4). 1000 repetitions were carried out for each combination of the parameters.

*matrix induced by the tree $T$. Note, when $p = \infty$, $L_\infty$ is the maximum difference between $D$ and $D_T$.*

As described in the previous section, *Treevolve* and the *phylip* package *dnadist* were used to generate the distance matrices for this simulation. Firstly, a neighbor-joining tree [84] was constructed from the simulated data, and the four measures of tree-likeness were recorded. The taxon $x$ with the highest $\overline{\delta}_x$ value was identified (ties were broken randomly). This taxon, $x$, was removed and a new neighbor-joining tree constructed on the reduced data set. In each experiment the five taxa with the highest $\overline{\delta}_x$ were sequentially removed, and the various measures of tree-likeness were recorded after each taxon had been removed. Simulations were carried out over the range of parameters shown in table 4.1.

Originally, the simulation was designed so that the removal order was determined by the $\overline{\delta}_x$ values in the complete data set. However, I found there was a greater improvement in the tree-likeness measures, when at each stage, the taxon with the highest $\overline{\delta}_x$ value was removed. The results, for sequence length $c = 500$, are shown in figure 4.9.

In each plot in figure 4.9 the top lines of each type (dotted and solid) represent the highest level of recombination, and the lowest lines represent no recombination. The dotted lines are the results for random removal order and the solid lines are the $\overline{\delta}_x$ based taxon removal order. The $L_1$ and $L_2$ norms have been normalised to account for the decreasing number of sequences. The normalisation factor was $(n^2 - n)^{1/p}$ where $p = 1$ or 2. No normalisation is relevant when $p = \infty$.

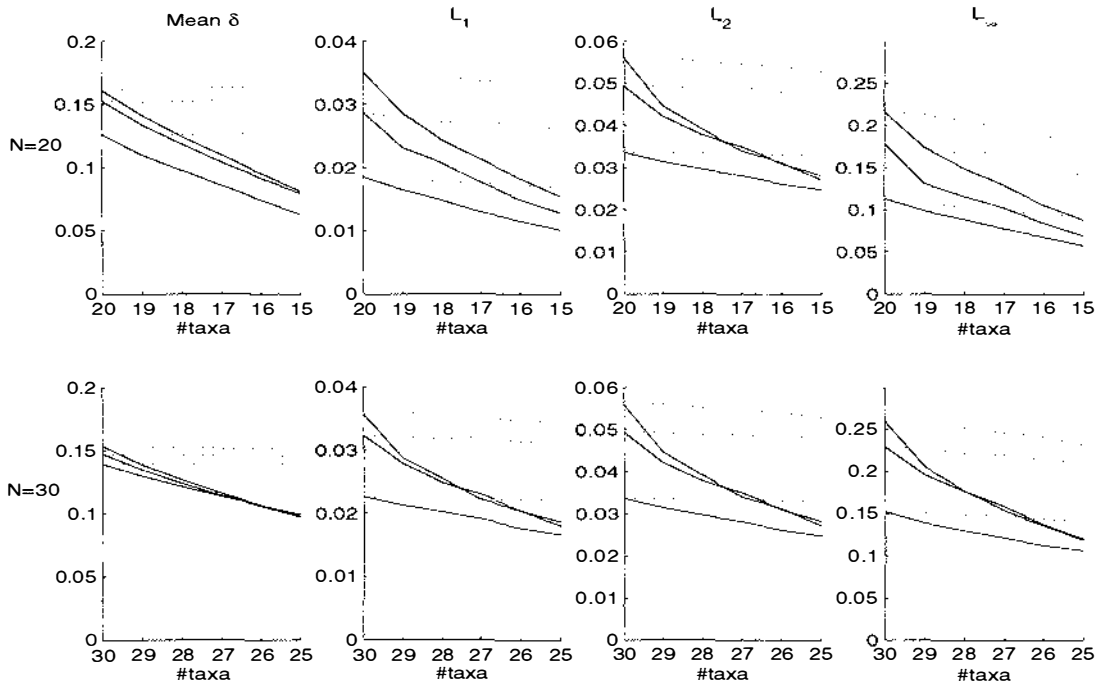


Figure 4.9: The effect of random versus $\overline{\delta}_x$ directed taxon removal orders on four measures of tree-likeness. Sequence length $c = 500$ in all plots. The top row of four plots is for $n = 20$ taxa and the bottom row is for $n = 30$ taxa. Within each plot, the top lines of each type (dotted and solid) represent the highest level of recombination $r = 1.0 \times 10^{-10}$, the middle lines are $r = 0.5 \times 10^{-10}$ and the lowest lines represent no recombination. The dotted lines are the results for random removal order and the solid lines are the $\overline{\delta}_x$-directed removal order.

The main trends displayed by the simulation results are:

1. The $\overline{\delta}_x$ based removal order is superior to the random removal order for all four measures of tree-likeness. The amount by which the $\delta$ directed removal order outperforms the random removal order is positively correlated with the amount of recombination.

2. With sequences of length $c = 100$ there is little difference between the plots of different recombination rates, this is because the recombination rate is defined per nucleotide, meaning that the longer sequences have a greater chance to recombine.

3. The removals cause a larger change in the measures of tree-likeness for 20 taxa data sets than for 30 taxa data sets. However, removing 2 taxa from the 20 taxa data sets (10%), gives a roughly equivalent improvement to removing 3 taxa from the data sets with 30 taxa.

4. The amount of recombination, $r$, is negatively correlated with all measures of tree-likeness.

In general the $\overline{\delta}_x$ based removal order leads to an improvement in the ability of neighbor-joining to construct a tree that represents the observed distance metric, whereas a random removal order gives little improvement.

### 4.5.2 Dependence of $\delta$ on topology

Before concluding that $\delta$-plots are a useful way of identifying potentially recombinant data sets, I wished to test the dependence of $\delta$ values on tree topology and edge lengths.

Using *Seq-Gen* version 1.22 [81], 1000 sets of sequences were evolved along two predefined trees. *Seq-Gen* was used instead of *Treevolve* because *Treevolve* doesn't allow you to evolve sequences along a specific tree. The first tree was the most

balanced topology on 16 taxa, and the second was the least balanced topology, sometimes known as the caterpillar tree (see figure 4.10). I expected that the plot of $\overline{\delta}_x$ for the balanced tree would be flat as all the taxa are identical to each other in terms of their position within the topology. However, with the caterpillar tree I expected that the quartets with many long branches would have higher $\delta$ values due to the increased probability of parallel changes and reversals. The $\delta$-plots for the balanced and caterpillar trees are shown in figure 4.11, and the average $\overline{\delta}_x$ values for taxa 1–16 are shown in figure 4.12.

The balanced tree has lower $\delta$ values on average ($\overline{\delta} = 0.0711$) than the caterpillar tree ($\overline{\delta} = 0.2502$). Taxa in the balanced tree are equally likely to have high mean $\delta$ values. However, with the caterpillar tree there are two noticeable trends. Firstly, as predicted, the taxa at the end of long edges have higher $\overline{\delta}_x$ values than those at the end of short edges. Secondly, the taxa in the middle of the caterpillar tree have higher $\overline{\delta}_x$ values. This may be because taxa in the middle of the tree are in many quartets that are nearly star-like, that is, the internal edge is small compared to the external edges. These plots show that the location of a taxon within the tree topology has an influence upon the $\delta$ values of the quartets to which it belongs. This is not unexpected, but it means we should be cautious before ascribing high $\delta$ values to recombination.

### 4.5.3 Identifying recombinant taxa

Sometimes recombination events occur that result in a set of taxa having one tree underlying some section of their aligned sequences, and a different tree underlying another section of the alignment. Frequently this can be described by some taxa changing their position within the tree, for two examples with viruses, see the Hepatitis B alignment discussed in [10], and the Dengue fever alignment of [47]. The simulation in section 4.5.1 shows that $\overline{\delta}_x$ can be used to identify those taxa that make it difficult to represent the observed metric as a tree, in this simulation

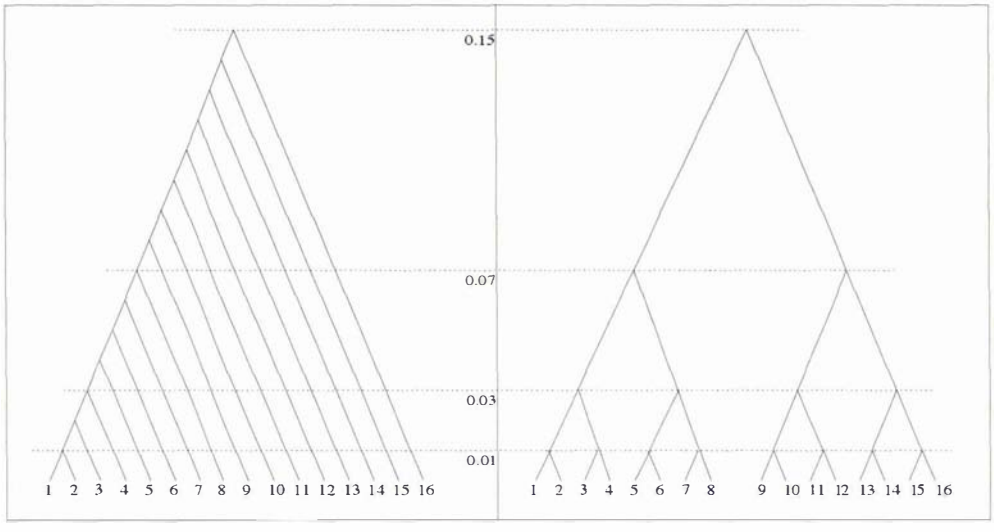


Figure 4.10: The two topologies used to generate the sequences. On the left is a caterpillar tree, it is the least balanced topology possible. The right hand tree is the most balanced topology. The expected number of changes from the root to the tips (0.15) is the same in both trees.

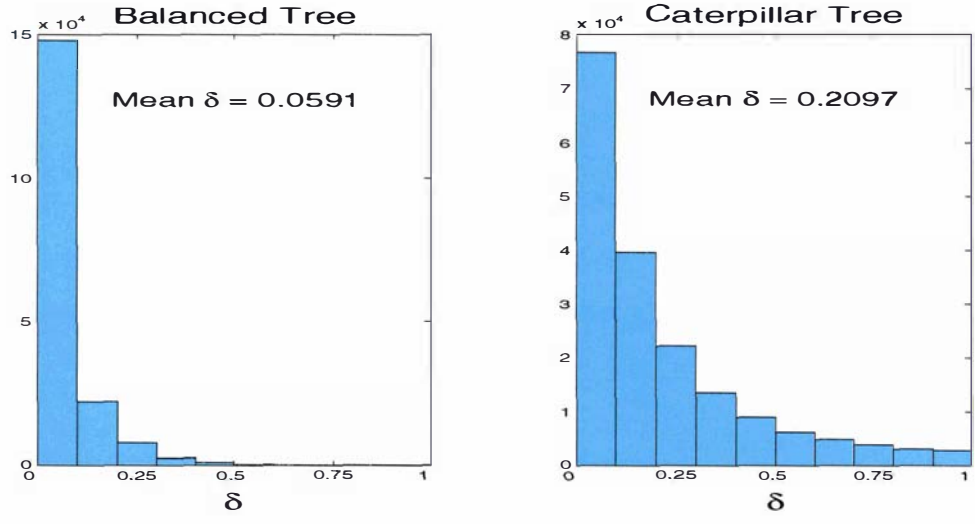


Figure 4.11: $\delta$-plots for the caterpillar and balanced trees. The plots represent averages over 1000 simulated data sets.

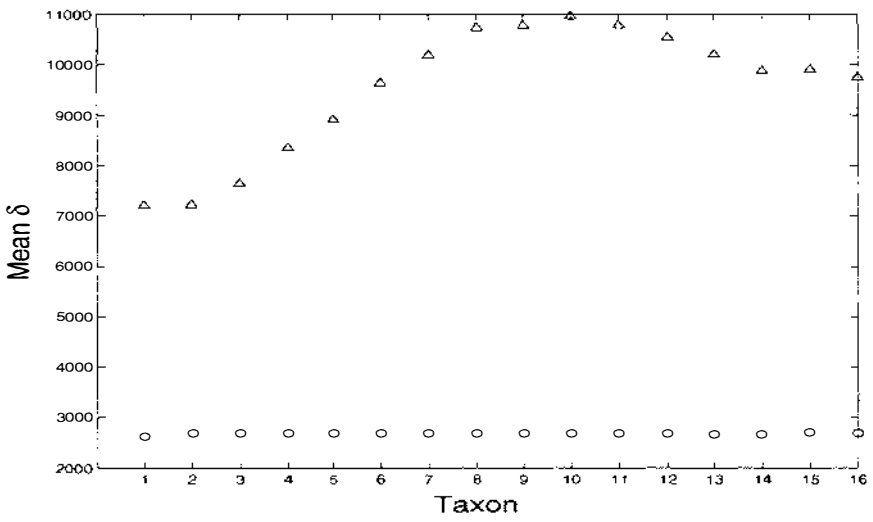


Figure 4.12: $\overline{\delta}_x$ averaged over 1000 simulated data sets. The + symbols are for the balanced topology, and the $\triangle$ symbols are for the caterpillar topology.

we wish to test if known recombinants can be detected via $\overline{\delta}_x$.

Figure 4.13 shows the trees used to generate the recombinant alignments used in the simulation. There were two basic topologies, either unbalanced (a) or balanced (b), each tree satisfies the molecular clock. With each basic topology the recombinant had parents that were either close (R1), intermediate (R2) or divergent (R3), giving $2 \times 3 = 6$ experiments in total.

Each sequence was 1000bp long, the sites 1-500 were simulated on the tree where the recombinant taxon (either R1, R2 or R3) was attached to its left-hand parent, sites 501-1000 were simulated along the tree where the recombinant taxon was attached to its right-hand parent. The model of sequence evolution used was HKY with $\kappa = 2$.

The results are shown in figure 4.14. It appears that it is easier to detect a recombinant sequence: (i) from within a balanced tree than an unbalanced tree; (ii) the more divergent the parents of the recombinant sequence.

I expected that shorter sequence lengths, and less symmetrical combinations of the "left" and "right" trees would make it harder to detect recombinant sequences.

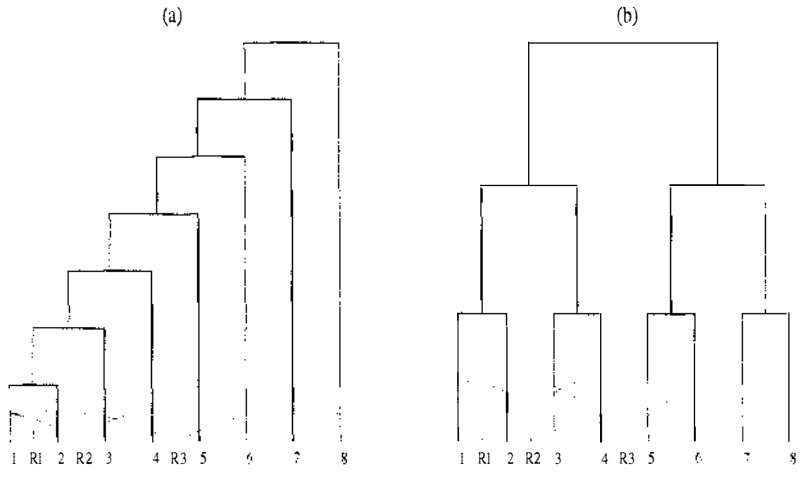


Figure 4.13: The trees used to generate the recombinant sequences. For example, with the unbalanced topology (a) and the close recombination event (R1) the first half of the sequences were generated on the tree where R1 and 1 are a neighbouring pair, and then concatenated with the sequences from the tree where R1 and 2 are a neighbouring pair.

To test this, the simulation was repeated with sequences of length 500 and 1,000, and left/right combinations of 50%/50%, 75%/25% and 90%/10%, giving $2 \times 3 = 6$ experiments for each of the six types of recombinant network described in figure 4.13. The results are only shown for the balanced tree where the recombinants parents are divergent (figure 4.15).

It should be easier to detect a single recombinant sequence from within a larger data set than a small one. This is because the ratio of quartets containing the recombinant, $\binom{n-1}{3}$, to the quartets containing a taxon $x$ and the recombinant, $\binom{n-2}{2}$, is $n : 3$. This ratio is only a rough guide to the ratio of the $\bar{\delta}_x$ values, as not all quartets containing a recombinant taxon will have high $\delta$ values. For example, consider the balanced topology shown in figure 4.13 (b), where the recombinants parents are close (R1). The quartet {1,5,8,R1} should not have a high $\delta$ value despite containing R1. This is because the topology of the tree on these four taxa does not change from one side of the breakpoint to the other, only the edge weights

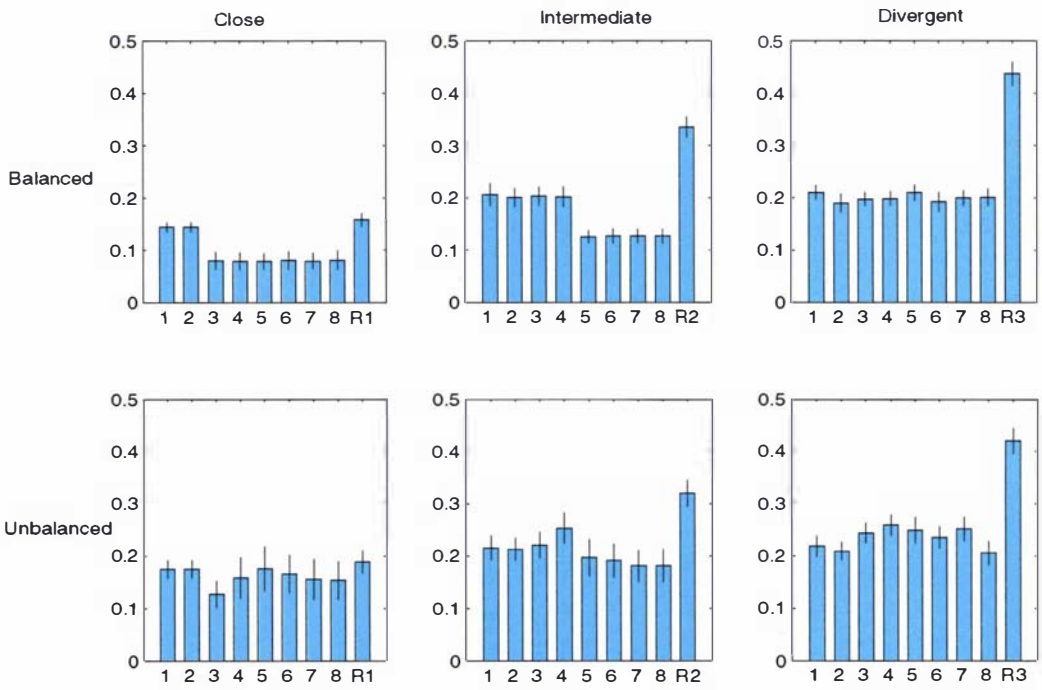


Figure 4.14: $\overline{\delta}_x$ for six types of recombinant alignment. The generating trees are shown in 4.13. In each plot the right-hand bar shows $\overline{\delta}_{R*}$ - the mean value of $\delta$ for the quartets containing the recombinant taxon. The vertical lines indicate $\pm$ one standard deviation.

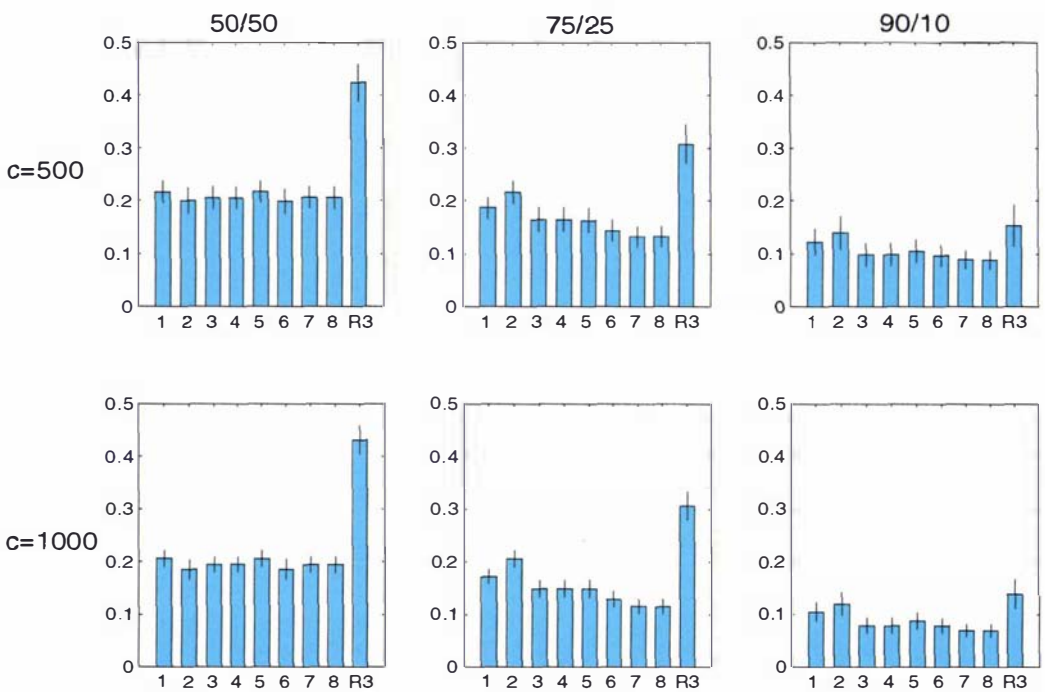


Figure 4.15: $\overline{\delta}_x$ for six combinations of sequence length and level of asymmetry. In each case the generating tree is the balanced topology with recombinant R3 (see figure 4.13). In each plot the right-hand bar shows $\overline{\delta}_{R3}$ - the mean value of $\delta$ for the quartets containing the recombinant taxon. The vertical lines indicate $\pm$ one standard deviation. In the top row of plots sequence length $c = 500$, and in the bottom row $c = 1000$. From left to right the combination of sequence from each parent of the recombinant is 50%/50%, 75%/25% and 90%/10%.

change. In fact, adding the two additive metrics for {1,5̄,8,R1} on either side of the breakpoint gives another additive metric. So, as $c \to \infty$, the expected value of $\delta$ for the quartet {1,5̄,8,R1} is $\mathrm{E}[\delta\{1, \bar{5}, 8, R1\}] = 0$.

This explains why, as is particularly noticeable for the simulations using the balanced tree, those taxa that are close to the recombinant, e.g. taxa 1 and 2 for R1, have higher values of $\overline{\delta}_x$ than the taxa further away. This is because the taxa that are far away, are in more quartets like the one described above, where the unweighted tree on the quartet does not change from one side of the breakpoint to the other.

## 4.6 Case Study: *Candida albicans*

In this section I use *Candida albicans* as a case study to test the methods discussed in this chapter. *Candida albicans*, a yeast, is an opportunistic pathogen that can cause disease in humans. Schmid et al. [86] presented evidence for a cluster of genetically similar isolates within *C. albicans* that is prevalent across many geographical regions, patient types, and forms of infection. A tree displaying this cluster is shown in chapter 5̄, figure 5̄.6, it will be referred to as cluster $S$.

*C. albicans* has been the subject of some debate recently over whether it is primarily a clonally or sexually reproducing organism. Different studies support conflicting results, see for example [38, 76, 101]. When a species reproduces clonally the isolates will be related to each other in a tree-like hierarchy, conversely if they reproduce sexually then one would not expect to see tree-like relationships between the isolates due to reassortment between the six chromosomes. $\delta$-plots were used as a visual tool to examine the tree-likeness of the isolates within and across cluster $S$.

Firstly, a distance matrix was constructed from 132 AFLP (Amplified restriction fragment length polymorphism) banding patterns on 42 isolates from the species *C. albicans*. 26 of these strains come from within cluster $S$, and 16 come from

outside the cluster. Each band in the AFLP data was originally coded as 0,1,2 or 3, depending on its strength. A binary character matrix was formed by mapping the data so that (0,1) $\rightarrow$ 0 and (2,3) $\rightarrow$ 1. Hamming distances were then taken between each pair of rows in the binary character matrix. The average pairwise distance between all isolates was 0.080, within cluster $S$ the average pairwise distance was 0.017, and between isolates not in cluster $S$ it was 0.095.

The $\delta$-plots in figure 4.16 show the marked difference in tree-likeness between cluster $S$ strains and the non-cluster strains. The non-cluster strains have $\overline{\delta} = 0.310$, compared to the cluster strains with $\overline{\delta} = 0.051$. One explanation for this could be that more recombination is occurring in the non-cluster strains, whereas the strains within the cluster are reproducing primarily clonally. Another possibility is that because the non-cluster strains are more diverse, the long edge lengths result in higher $\delta$ values.

The original data set analysed by Schmid et al. [86] was from RFLP on 266 isolates. For comparison to the $\delta$-plots from AFLP data, I have shown the $\delta$-plots for the original RFLP data set in figure 4.17. The number of samples was large ($n = 266$) so instead of computing $\delta$ for all quartets, 100,000 were chosen at random. In these plots $\overline{\delta} = 0.429$ for the cluster ($n = 98$), and 0.420 for its complement ($n = 168$); in each case, close to the value of $\sim$0.5 observed for the random data in figure 4.3. The higher $\overline{\delta}$ values for the RFLP data (0.429 and 0.420) compared to the AFLP data (0.051 and 0.310) highlights the fact that we should be cautious before attributing high values of $\overline{\delta}$ to recombination, as it may be that the method of measuring the distances adds a large amount of noise.

The standard technique used to test if a population is sexual or asexual is linkage analysis [108]. The idea is that factors linked along a chromosome of an organism will remain linked in its offspring if it reproduces clonally. However, if reproduction is sexual then reassortment will cause the factors to become unlinked. All pairs of columns in the AFLP data are tested to see if they are in equilibrium, or in statistical terms, if they are independent. Linkage analysis was performed using

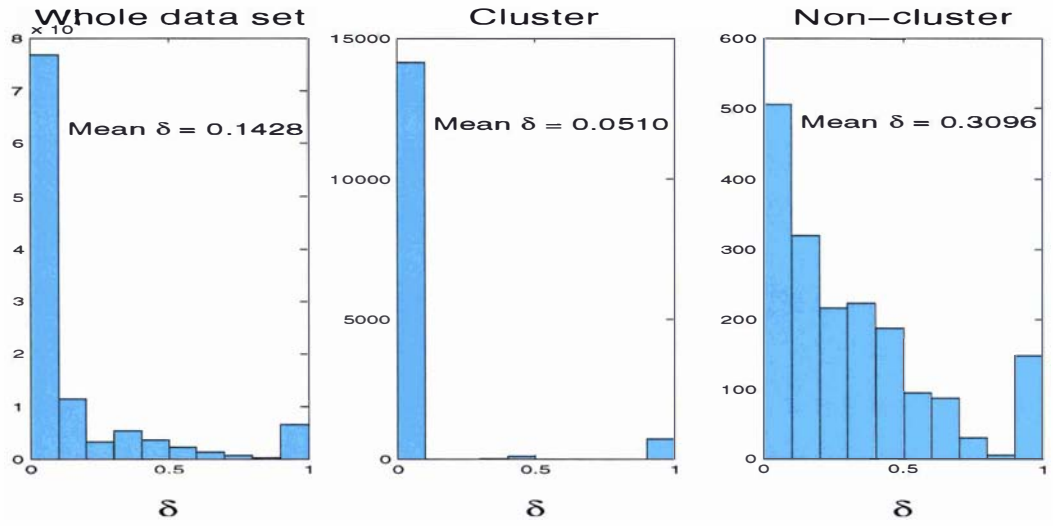


Figure 4.16: The $\delta$-plots for the whole AFLP data set ($n = 42$), the cluster ($n = 26$), and the non-cluster strains ($n = 16$).

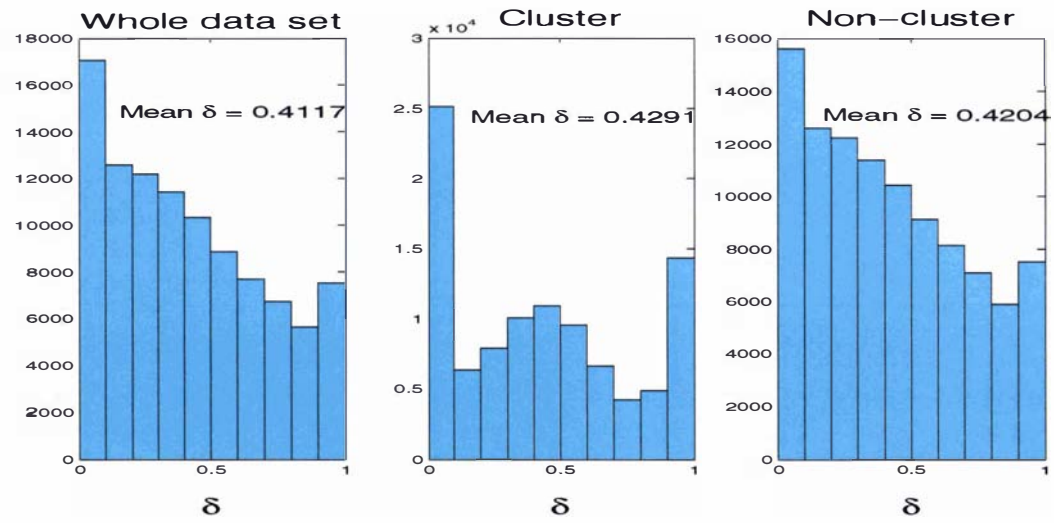


Figure 4.17: The $\delta$-plots for the whole RFLP data set ($n = 266$), the cluster ($n = 98$), and the non-cluster strains ($n = 168$). Each histogram represents the $\delta$ values of 100,000 randomly chosen quartets.

the exact test for Hardy-Weinburg equilibrium (HWE).

The exact test for disequilibrium for two-state alleles works as follows: Given two columns showing the presence or absence of some feature, i.e. a pair of bands in an AFLP, test the null hypothesis that the columns are independent. The two columns are summarised in a $2 \times 2$ table counting the number of times each pattern of presence (+) and absence (−) occurs.

| | + | − | Total |
|---|---|---|---|
| + | $n_{++}$ | $n_{+-}$ | $n_{+\cdot}$ |
| − | $n_{-+}$ | $n_{--}$ | $n_{-\cdot}$ |
| Total | $n_{\cdot+}$ | $n_{\cdot-}$ | $n$ |

Constant columns, where all isolates have the same value, are excluded from the analysis. For each possible two by two table with the observed marginal totals ($n_{+\cdot}$, $n_{-\cdot}$, $n_{\cdot+}$ and $n_{\cdot-}$), the probability of that pattern occurring is calculated as:

$$Pr(n_{++}, n_{+-}, n_{-+}, n_{--} | n_{+\cdot}, n_{-\cdot}) = \frac{n_{+\cdot}! n_{-\cdot}! n_{\cdot+}! n_{\cdot-}!}{n_{++}! n_{+-}! n_{-+}! n_{--}! (2n)!}$$

The possible patterns are ranked in order of their probabilities, and the cumulative probability distribution is calculated. If the observed pattern occurs at less than 5% in the cumulative distribution then the null hypothesis of independence is rejected. For a detailed explanation of this test see Weir [108].

The results for the AFLP data are summarised in table 4.2. The linkage analysis supports the same conclusion as the $\delta$-plots. Within the cluster, equilibrium could be rejected at the 5% level for 27% of the pairs of columns (46 out of 171). For the non-cluster strains about 11% (57 out of 528) of pairs of columns were rejected at the 5% level.

The HWE test is not effective when the number of two by two tables with the observed marginal sums is small. For example, if $n_{+-} = n_{-+} = 0$, $n_{--} = 1$, then there are no other matrices with the same marginal sums, and the $p$-value returned

| | Variable columns | Rejects at 5% |
|---|---|---|
| Whole data set | 40 | $\frac{270}{780} = 34.6\%$ |
| Cluster | 19 | $\frac{46}{171} = 26.9\%$ |
| Non-cluster | 33 | $\frac{57}{528} = 10.8\%$ |

Table 4.2: Summary of the linkage analysis for *Candida albicans*. Pairs of variable columns were tested for linkage equilibrium. The table shows the number of non-constant columns and the proportion of rejections at the 5% level.

by the HWE test is 1. This doesn't seem to be an appropriate test, as the columns in the binary character matrix are compatible and could be considered tree-like. The extent of this problem is highlighted in figure 4.18 showing the distribution of $p$-values. I suggest that pairs of columns with two marginal sums equal to one should not be considered in the analysis, as they are non-informative. If all pairs of this type are removed from the analysis, the proportion of rejections for the cluster strains becomes $\frac{46}{93} = 49.5\%$, the proportion of rejections for the non-cluster strains is not changed.

Both the $\delta$-plot method, and the traditional method of linkage analysis, support the claim that cluster $S$ is primarily clonal, in contrast to the non-cluster strains. It would be interesting to investigate this claim experimentally.

### 4.6.1 $\overline{\delta}_x$ for *C. Albicans* data

Given that the $\delta$-plots for the AFLP data suggest that cluster $S$ is much more tree-like than its complement, I expected that in a plot of $\overline{\delta}_x$ for the combined data set the non-cluster strains would have significantly higher values. This did not turn out to be the case, perhaps due to the strong split in the data between the cluster and non-cluster strains combined with the differing numbers of isolates in each category. The table 4.3 shows that the majority of the non tree-like quartets occur when three isolates come from within cluster $S$ and one from without, or vice versa. There are more opportunities for the cluster strains to be in "bad" quartets than there are for the non-cluster strains. This example illustrates that tree topology

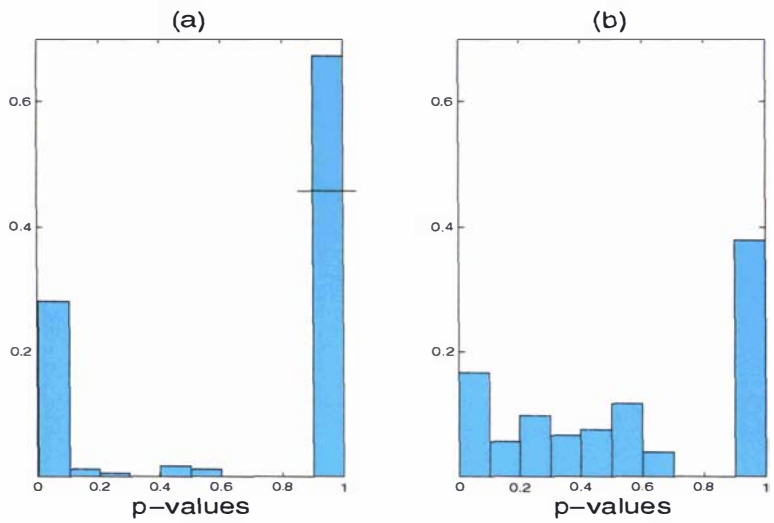


Figure 4.18: The $p$-value distribution for the cluster strains(a) and the non-cluster strains (b). The first bar on the left hand side of each plot is the 5% rejection zone. The majority of the $p$-values of 1 in (a) result from the situation described in the text where two of the marginal sums are 1. The proportion of such $p$-values are those below the line crossing the right-hand bar.

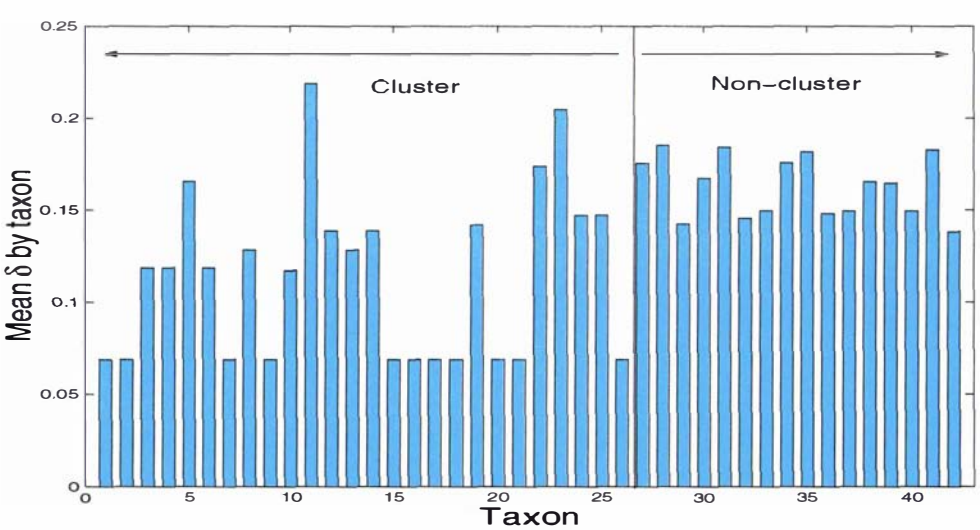


Figure 4.19: Plot of $\overline{\delta}_x$ for the *C. albicans* AFLP data set. The counts for the cluster (taxa 1–26) are frequently as large as for the non-cluster taxa (27–42).

| | 4C/0NC | 3C/1NC | 2C/2NC | 1C/3NC | 0C/4NC |
|---|---|---|---|---|---|
| #Quartets | 14,950 | 41,600 | 39,000 | 14,560 | 1,820 |
| $\overline{\delta}$ | 0.0510 | 0.1556 | 0.0865 | 0.3307 | 0.3096 |

Table 4.3: In this table all the possible quartets are divided into five categories. For example, 4C/0NC means that all four taxa in each of the quartets are from within cluster $S$, 2C/2NC means that 2 taxa in each quartet are from the cluster and two are not within the cluster. The first row of the table shows the number of quartets in each category and the second row shows $\overline{\delta}$ for that category.

has a strong influence on $\delta$ values.

## 4.7 Discussion

$\delta$-plots are a visual extension of statistical geometry that provide an initial analysis of phylogenetic data sets. They measure how tree-like a data set is, and hence give an indication of how much confidence may be placed in the output of tree estimation algorithms. They could be used as an alternative to, or in combination with, linkage analysis as a tool for identifying recombination in a data set. The main limitation is that a large number of factors can influence $\delta$ values, such as random noise (e.g. from short sequence length), the method used to infer distances, reticulation, and the underlying tree topology. Hence, it is difficult to use $\delta$-plots as a diagnostic tool to identify a single feature of the data such as the presence of recombination.

The quantity $\overline{\delta}_x$ was used to identify those taxa which most confounded the tree-like signal. The removal of these taxa resulted in an improvement in neighbor-joinings ability to represent the observed metric. Furthermore, $\overline{\delta}_x$ was used to identify recombinant taxa in an experiment with simulated data. The position of a taxon within the tree topology was found to influence the $\delta$ values of the quartets containing that taxon.

**Directions for future research**

- Measure $\overline{\delta}$ within a sliding window to investigate if it is a good indicator of recombination breakpoints.
- Test if $\overline{\delta}$ is a effective measure for determining the most appropriate distance correction.

# Chapter 5

# Selecting Good Model Strains: Examples from Microbiological Research

## 5.1 Introduction

This chapter describes work I did in collaboration with Massey microbiologist Jan Schmid. The general task was to develop a quantitative approach to selecting representative model strains. The examples used are microbiological pathogens, but the problem of selecting model strains occurs more generally within biology.

A model strain is an individual chosen to represent a species or sub-species. Once selected, model strains are studied in greater depth than the other isolates, for instance, they may be partially or, in the case of genome projects, completely sequenced. Furthermore, model strains for pathogenic species are often used to develop new drugs. To be a good model strain an isolate must be representative of its species. It would, for example, be a huge expense to develop drugs that were only effective against a small proportion of a pathogenic species. It may be that the species in question is subdivided into clusters or subspecies with different properties. In these cases it would be advantageous to have model strains from

each cluster. We seek the answers to two interrelated questions:

- What natural clusters, or sub-species, are there within the data?
- Which isolates are the most representative model strains within these clusters?

The specific application, of most interest to Jan Schmid, was to choose good model strains for the pathogenic yeast species *Candida albicans*. This formed a part of an ongoing project studying this species. Throughout this chapter *C. albicans* is used as an example along with two bacterial species *Helicobacter pylori*, and *Pseudomonas aeruginosa*.

In the following section (5.2) I seek to show why choosing representative model strains is an important task. I also discuss the role of tree construction in identifying clusters within the data. In section 5.3 methods are developed for identifying clusters and choosing model strains. These methods fall into three groups, the first group of methods are based directly on dissimilarity information, the second analyses quartets in the data, and the third is an adaption of the graph theory problem dominating set [36]. Some artificially constructed examples are used throughout this section to illustrate different properties of the methods. In section 5.4 trees are constructed for the three microbiological data sets: *C. albicans*, *H. pylori*, and *P. aeruginosa* to identify major clusters. The methods developed for choosing model strains are applied to these three data sets, and the results are compared.

## 5.2 Motivation

Many types of biological research, such as whole genome sequencing, the search for virulence factors and drug targets in microbiological organisms, or phylogenetic analysis of the relationships between species, rely on the use of model strains purported to represent a species under study. In some cases a single model strain can be chosen from a collection of isolates on the basis that it is typical of, or in some

sense central to, the other isolates. However, if the species is subdivided into major sub-species, a single strain chosen because it is central, might not be representative of any of the sub-species, but instead a rare type. In this case it may be preferable to use several model strains, one per sub-species. Alternatively, if financial considerations mean that only a limited number of strains can be used, the best choice of model strains might be those that represent the predominant sub-species.

The topic of model strain selection does not appear to have been addressed in the literature, so it is unclear how they are currently selected. It may be that most model strains are chosen for reasons of expediency, as in the *P. aeruginosa* genome project [94]. In this paper the authors state:

> "Strain PA01, a wound isolate, was chosen as a strain prototype for sequencing because it is the most widely used *P. aeruginosa* laboratory strain and because physical and genetic maps were available."

While these are good reasons, they do not guarantee that PA01 will be a representative example of *P. aeruginosa*.

Each of the three example data sets *C. albicans*, *P. aeruginosa*, and *H. pylori* cause disease in humans [51, 70, 99]. They are each currently the subject of genome sequencing projects[1] [2, 94, 102], and there are efforts towards developing effective drug treatments. Hence, it is important that the model strains chosen from these species are representative.

As stated in the introduction, one aspect of choosing good model strains is identifying any major clusters (sub-species) that occur in the data. While the notion of a cluster (or group) is widely used, there is no agreed upon rigorous definition of what a cluster is. Loosely speaking, we expect that members of a cluster are similar to each other, and also that they are dissimilar to the rest of

[1] See, **http://www.cmcb.uq.edu.au/aeruginosa/summary.html**, **http://gib.genes.nig.ac.jp/Hp99/top.html**, and **http://www-sequence.stanford.edu/group/candida/** for links to the three genome projects.

the set of interest. These two properties have been termed *internal cohesion* and *external isolation* [30, 37].

In species that reproduce by binary fission, such as the bacteria *H. pylori* and *P. aeruginosa*, it is reasonable to expect isolates to be related in a tree-like fashion, because there is no recombination of the genetic material (although, small amounts of DNA in plasmids can be transferred between isolates [51, 99]). This makes constructing a tree an obvious first step in determining clusters. A rooted tree can be considered as a hierarchy of clusters, where each clade in the tree corresponds to a cluster. For each pair of clusters, either one cluster is a subset of the other, or they have no isolates in common. That is, given two clusters $C_1$ and $C_2$ in a hierarchy, their intersection $C_1 \cap C_2$ must be in the set $\{C_1, C_2, \emptyset\}$.

In an unrooted tree, the removal of an edge will split the data into two clusters. How useful these clusters are for helping to determine model strains, will depend on the sizes of the two subsets and the length of the edge. For the best example of a tree showing obvious clusters in this thesis, see the tree of the Adélie penguin data shown in chapter 3, figure 3.5, the removal of the central edge splits the data into the Antarctic cluster and the Ross Sea cluster.

In contrast to the previous chapters, where trees are constructed to reflect the genealogy of a set of taxa, when using trees to identify clusters, the genealogy is unimportant. Indeed, in cases where isolates undergo sexual reproduction, a tree will not be an appropriate reflection of the genealogy, which will be reticulate. What is important is that the isolates grouped together behave in a similar way. So, for instance, if a model strain is chosen to represent one cluster, then the drugs developed to treat that strain should have a similar effect on other isolates within the cluster. As we can't measure behaviour directly, we hope that the dissimilarity information is an adequate enough reflection of it that the clusters found are biologically meaningful.

## 5.3 Methods

Firstly, I introduce some notation. Let $X$ be a set of $n$ isolates, and $M = \{m_1, m_2, \ldots, m_k\}$ a set of $k$ model strains, where $M \subset X$. For all $x, y \in X$, $d(x, y)$ measures the dissimilarity of $x$ and $y$.

DEFINITION: 7 (DISSIMILARITY) *A dissimilarity* $D : X \times X \rightarrow \mathbb{R}$ *satisfies the conditions:*

1. $d(x, y) \geq 0$, $d(x, x) = 0$, *for all* $x, y$

2. $d(x, y) = d(y, x)$, *for all* $x, y$

Most biological data satisfies these properties, and frequently the triangle inequality also (see chapter 1, section 1.3). The data in this chapter is derived from (restriction fragment length polymorphism) RFLP, and is always symmetric and non-negative, an example data set in section 5.4 illustrates why this is so, and also why the triangle inequality might not be satisfied.

An ideal method for selecting model strains would take as input a dissimilarity matrix on a set of taxa, and output both which, and how many model strains are "optimal". However, to compute some measure of optimality for all possible sets of model strains is computationally intractable for all but the smallest data sets. For a set $X$ of $n$ isolates, the number of possible sets of model strains $M$, where $M \subset X$, is $2^n - 1$ (the power set of $X$ minus the empty set).

Given unlimited time and money, it would always be optimal to have $n$ model strains, in other words, each isolate representing itself. In practice the number of model strains, $k$, will usually be determined in advance, or fixed within a tight range, due to financial or time considerations. For instance, in many genome projects only a single strain is sequenced. For a predetermined $k$ there are $\binom{n}{k}$ possible sets of model strains. Once an optimal set of $k$ model strains is found, this defines a set of $k$ disjoint subsets of $X$. These subsets (clusters) are determined by first initialising each subset to contain an individual model strain, and then

assigning each element in $X$ to the subset containing the model strain it is closest to (with ties broken randomly).

More restrictively, if the clusters in the data have been predefined, perhaps on the basis of major features of the tree, then it is computationally cheap to choose one model strain per cluster. All of the methods described below, with the exception of the quartet based method which only works with predefined clusters, apply to both the case where $k$ is decided, and the case where the clusters have been predetermined and one model strain per cluster is required.

### 5.3.1 Dissimilarity Based Methods

Firstly, I discuss the case where clusters have been predetermined and the aim is to select one model strain per cluster. One criteria is, for each cluster $C \subset X$, to choose the strain $m \in C$ that has the smallest sum of dissimilarities to the other strains in the cluster

$$\min_{m \in C} \sum_{i \in C} D(m, i). \tag{5.1}$$

Alternative criteria are to minimise the sum of squares,

$$\min_{m \in C} \sum_{i \in C} D(m, i)^2, \tag{5.2}$$

or, minimise the maximum dissimilarity to any isolate,

$$\min_{m \in C} \max_{i \in C} D(m, i). \tag{5.3}$$

More generally, if the clusters have not been specified, we can search for the optimal set of $k$ model strains. To select $k$ model strains choose $M = \{m_1, m_2, \ldots, m_k\} \subset X$ such that the cost $C_k(M)$ is minimum. The generalisa-

tion for equation 5.1 is

$$C_k(M) = \sum_{i \in X} \min_{m \in M} D(m, i) \tag{DC1}$$

And for equations 5.2 and 5.3 respectively

$$C_k(M) = \sum_{i \in X} \min_{m \in M} D(m, i)^2 \tag{DC2}$$

and

$$C_k(M) = \max_{i \in X} \min_{m \in M} D(m, i) \tag{DC3}$$

(These criteria are henceforth referred to as DC1, DC2 and DC3.)

Ideally, we would like to extend these criteria to give an optimal number of model strains. One strategy would be to compare $C_k(M)$ for each possible set of model strains $M$, and $k \in \{1, 2, \ldots, n\}$. As has already been discussed, there are two main problems with this approach. The first, is combinatorial explosion, there are $2^n - 1$ putative sets of model strains that need to be checked. Secondly, it will always be optimal to have $n$ model strains, in other words each isolate representing itself. In general, if $M_k^*$ is a minimum cost set of $k$ model strains then $C(M_k^*) \geq C(M_{k+1}^*)$. To see this is true, consider $M_k^*$, an optimal set of $k$ model strains. Choose any strain $m \in X \backslash M_k^*$. Set $M_{k+1} = M_k^* \cup m$. It is easily seen that $C(M_k^*) \geq C(M_{k+1}) \geq C(M_{k+1}^*)$. Hence, some procedure would be needed to test the significance of $C(M_k^*) - C(M_{k+1}^*)$. This idea is not explored further in this chapter, but see Everitt [30], pg 100, for a discussion of the similar problem of deciding on an optimal number of clusters.

Our aim is to choose representative model strains, but what does the word "representative" mean in mathematical terms? The three criteria above, minimising the sum of distances, minimising the sum of distances squared, or minimising

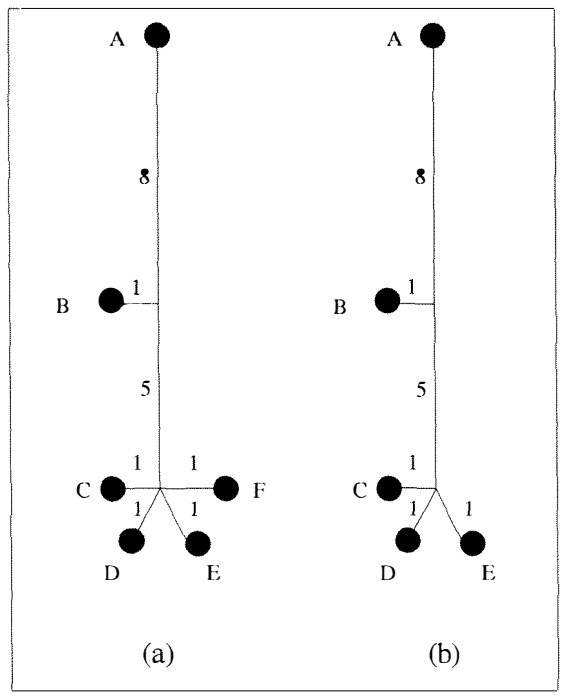


Figure 5.1: In panel (a) using criteria DC1 or DC2 any of C, D, E or F are a minimum cost choice of model strain ($k = 1$). Using criteria DC3, B is the best model strain. In panel (b) F has been removed. In this case criterion DC1 still picks C, D or E as the best model strain, but now both criteria DC2 and DC3 pick B as the best model strain.

---

the maximum distance, of all strains to their closest model strain, differ in their interpretation of "representative". While there is no objective way of deciding which of these criteria is best, we can study how they behave on test cases and judge which criteria might be most appropriate for a particular application.

Figure 5.1 illustrates a situation where the three criteria pick different model strains. The example is an additive metric but the principle applies to dissimilarities in general. Criteria DC1, DC2 and DC3 place progressively more importance on not being a large distance from any strain, in exchange for being a moderate distance from many strains.

What is the computational complexity of this group of methods? There are $\binom{n}{k}$ ways of choosing $k$ strains from $n$ isolates, and for each set $M$ of model strains it requires $nk$ operations to calculate $C_k(M)$, because for each of the $n$ isolates,

$k$ putative model strains are checked to determine which it is closest to. So, the computational complexity of choosing strains according to these criteria, for fixed $k$, is $O(n^{k+1})$.

### 5.3.2 Quartet Based Method

Given a postulated cluster $C \subset X$, if you form a quartet from two elements $a, b \in C$, and two elements $a', b' \in X \backslash C$, then you would expect $d(a,b) + d(a',b') < \min(d(a,b') + d(a',b), d(a,a') + d(b,b'))$. If $D$ is an additive metric, and the clusters $C$ and $X \backslash C$ are defined by removing an edge in the unique tree corresponding to this metric, then the first sum being smallest follows from the four point condition. Even for non-treelike data it should follow from the intuitive definition of a cluster having *internal cohesion* and *external isolation* [37].

In Schmid et al. [86] this idea is used to measure how strongly their data supports specific edges that define clusters of interest. For each edge defining a cluster $C$, they report the proportion of times that for a randomly chosen quartet with two elements $a, b \in C$, and two elements $a', b' \in X \backslash C$, the sum $d(a,b) + d(a',b')$ is the smallest of the three sums above. Schmid et al. argue that for a randomly selected subset of $X$, this proportion will be $\frac{1}{3}$ (due to symmetry), and they report that the cluster $S$ they define (see figure 5.6) is significant, as the proportion is $\frac{23}{30}$, much greater than $\frac{1}{3}$.

I would further note, that for many types of data generated on a non-tree model (for example, road map distances), if you estimate a tree for the data, and then calculate the average proportion of quartets that support an edge in the tree, you would expect the proportion to exceed $\frac{1}{3}$. To expect a proportion of $\frac{1}{3}$ the tree would have to be constructed randomly as well. Instead, trees are constructed on the basis of similarity, and even for random data it is possible to group similar things. Hence, it does not follow that the cluster corresponds to a clade from an underlying tree-like process.

While the procedure of Schmid et al. [86] may indicate some measure of the support for an edge it doesn't identify which isolates to choose as model strains. I extended the idea as follows: Given a cluster $C \subset X$ and an element $m \in C$. Let $Q_m$ be the set of all quartets $q$, containing the fixed element $m$, of the form $q = \{m, a, b, c\}$ where $a \in C$, and $b, c \in X \backslash C$. If

$$d(m,a) + d(b,c) \leq d(m,b) + d(a,c)$$

and

$$d(m,a) + d(b,c) \leq d(m,c) + d(a,b)$$

then we say that $q$ **supports** $C$. Let $y(q)$ be an indicator variable such that

$$y(q) = \begin{cases} 1, & \text{if } q \text{ supports } C \\ 0, & \text{otherwise} \end{cases}$$

The quartet based criterion (QBC) chooses as a model strain, for cluster $C$, the isolate $m \in C$ that is in the largest number of quartets that support $C$

$$\max_{m \in C} \sum_{q \in Q_m} y(q) \qquad \text{(QBC)}$$

The computational cost of choosing the best model strain by this method is proportional to the number of quartets with two elements in the cluster $C$, and two in its complement $C' = X \backslash C$, $\binom{n_C}{2} \times \binom{n_{C'}}{2}$, where $n_C$ is the number of elements in $C$, and $n_{C'}$ the number of elements in $C'$. So, the computational complexity of the QBC is $O(n^4)$.

Figure 5.2 illustrates a weakness of this method. As an example, consider a data set with an additive metric, and two clusters that correspond to a split in the tree defined by the additive metric (figure 5.2 (a)). All quartets in $Q_m$ obey the four-point condition, and therefore support $C$, for every choice of $m$. Hence, all

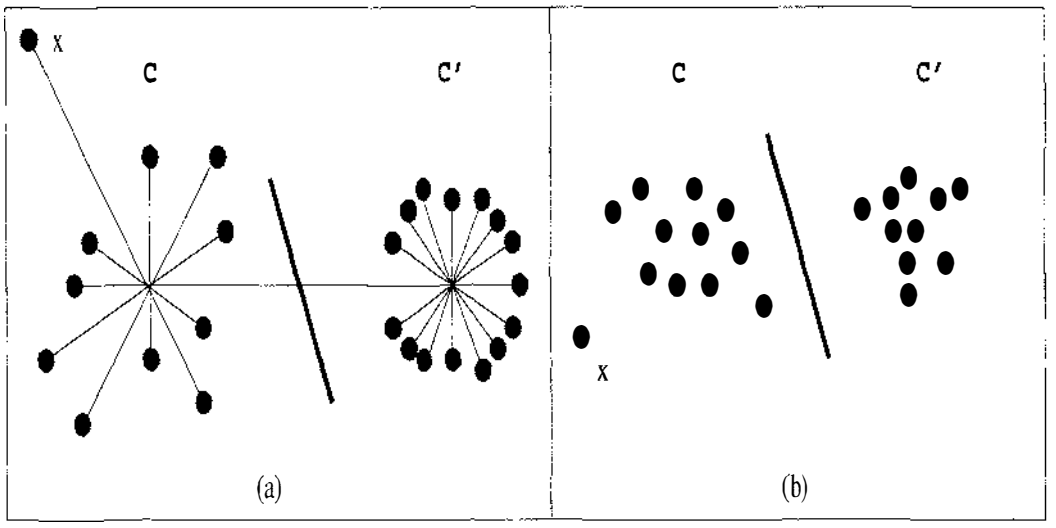


Figure 5.2: On the additive metric generated by the tree in panel (a), all isolates in cluster $C$ are considered to be equally good model strains for $C$ under the QBC. Similarly in the non-additive situation in panel (b), for points scattered in the Euclidean plane, the QBC may consider isolate $x$ to be a good model strain for cluster $C$.

isolates will be considered equal by the QBC. However, some isolates may be very distant from the other strains in the cluster to which they belong. Figure 5.2 (b) shows a similar example with non-additive data.

### 5.3.3 Graph theoretic approach

In the graph theory problem **dominating set** [36] a problem instance consists of a graph $G = (X, E)$, where $X$ is the node set and $E$ is the edge set, and an integer $k < |X|$. The question is:

*Does there exist a subset* $M \subset X$ *of size* $k$ *or less such that for all* $u \in X \backslash M$ *there exists an element* $v \in M$, *for which* $\{u, v\} \in E$.

In order to express the problem of selecting model strains in terms of dominating set, we first choose a threshold value, $T$. This threshold is the maximum allowable dissimilarity between any isolate and a model strain, for which that model strain is considered to be an adequate representative of the isolate. Isolates are represented by nodes in the graph $G = (X, E_T)$, and a pair of isolates $x, y \in X$ are connected

by an edge $e \in E_T$ provided that their dissimilarity is less than the threshold $T$.

$$D(x, y) < T \iff \{x, y\} \in E_T$$

The question,

> "Is there a dominating set of size $k$ for the graph $G = (X, E_T)$?"

is equivalent to asking,

> "Do there exist model strains such that every isolate is less than a dissimilarity of $T$ from at least one model strain?"

Note, the problem could also be phrased in an optimisation framework where the question becomes: *What is the* **least** $k$ *such that there exists a set* $M$ *of size* $k$ *where* $M$ *is a dominating set of the graph* $G$.

In many biological applications, budget constraints on a project mean that the number of affordable model strains is specified in advance. Simply knowing whether or not a dominating set exists, for a given $T$ and $k$, may not be useful. The following criteria for selecting model strains is based on the dominating set problem, but allows different sets of model strains to be ranked in order of how representative they are. Given a set of model strains $M$ and a threshold $T$, there will be some isolates $x \in X$ where $d(m, x) \leq T$ for some $m \in M$, these are said to be **reachable**, and others where $d(m, x) > T$ for all $m \in M$, that are not reachable. One set of model strains is said to be more representative than another if it **reaches** more isolates.

More formally, for a set of model strains $M$, and a threshold $T$, the number of reachable isolates is $r$ where

$$r = r(M, T) = \sum_{x \in X} y(x)$$

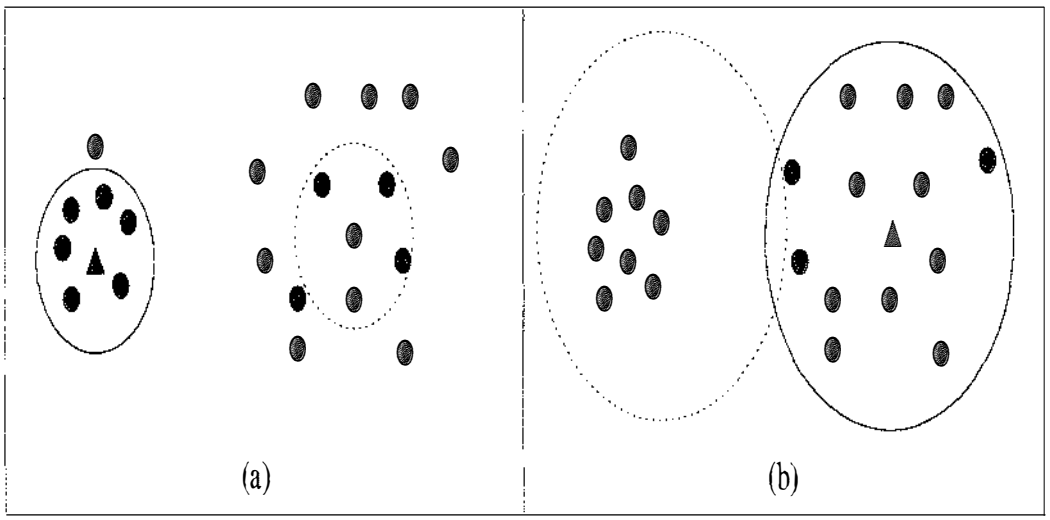


Figure 5.3: An example data set in the Euclidean plane, with two obvious clusters. With a small value of $T$, as in panel (a), and $k = 1$, the triangle shows the isolate that has been chosen as a model strain. The circle indicates which isolates are within the threshold value $T$. The number of isolates reached is $r^* = 7$. If a second model strain was chosen it would come from the right-hand cluster and reach the five isolates indicated by the dotted circle. In panel (b) $T$ is larger, so the model strain instead comes from the right-hand cluster which contains more isolates, although the within cluster distances tend to be larger.

with

$$y(x) = \begin{cases} 1, & \text{if } D(x,m) \leq T \text{ for some } m \in M \\ 0, & \text{otherwise} \end{cases}$$

Let $r^* = r^*(T)$ be the maximum number of isolates reachable by any set of $k$ model strains given a threshold value of $T$. The dominating set based criterion (DSC) chooses a set of model strains $M$, that reaches $r^*$ isolates. An example of the DSC is shown in figure 5.3.

Dominating set is an NP-complete problem [36]. There is no known algorithm for answering the above questions other than exact search, that is, checking all $\binom{n}{k}$ sets of size $k$. The above criterion for selecting model strains is clearly equivalent in computational complexity to the dominating set problem. If a set $M$ that gives the maximum value of $r$ can be determined in polynomial time, and it is $O(1)$ to check if $r = n$, then it would take polynomial time to check if there is a dominating

set.

### 5.3.4 Greedy algorithms

For large data sets it is too time intensive to check every possible set of model strains to find those that are optimal according to some cost function. For the criteria DC1, DC2, DC3 and DSC, greedy approximations to the exact search algorithms can be used. Model strains are chosen one at a time, at each stage the isolate that gives the best improvement in the cost function is chosen as a model strain.

Pseudocode is given for a greedy algorithm for the dissimilarity criteria. The algorithm takes as input the dissimilarity matrix $D$, and the specified number of model strains $k$, and returns `bestCost`, the value of $C_k(M)$ for the greedy choice of model strains `bestM`.

```
(bestCost, bestM) = greedy(D,k)


M = ∅;
bestCost = LARGE; An arbitrary large constant.
bestM = ∅;
Repeat loop to choose each of k model strains.
for loop = 1 to k {
      for i = 1 to n { Check all isolates (n=#X),
            if i ∉ M { that are not already in M.
                  cost = f(M ∪ i, D); f could be DC1, DC2 etc.
                  if cost < bestCost {
                        bestCost = cost;
                        bestM = M ∪ i;
                   }
             }
       }
M = bestM; Update M.
}
```

The computational complexity of the greedy algorithm is $O(n)$ for fixed $k$, because for each of $k$ model strains, $n$ candidates must be checked to see if they are the best choice for inclusion into the model strain set.

The difference in running time for the exact and greedy versions of D1 for $k = 1, 2, 3, 4$ is shown below. The exact value for $k = 4$ has been extrapolated from the observed times for $k \leq 3$, based on the number of operations it would have to perform.

| k | 1 | 2 | 3 | 4 |
|---|---|---|---|---|
| Exact | $<$ 1 sec | $\sim$2 min | $\sim$4.5 hrs | $\sim$300 hrs |
| Greedy | $<$ 1 sec | $<$ 1 sec | $<$ 2 sec | $<$ 3 sec |

The Greedy algorithm is particularly useful for this problem application, where the detrimental effect of producing a good, but not optimal, solution is insignificant. What needs to be avoided is choosing terrible model strains, as the financial expense of choosing a strain that had very different properties to the norm would be high.

## 5.4 Analysis of example data sets

The dissimilarity matrices for the three example data sets, *P. aeruginosa* , *H. pylori* and *C. albicans*, are all from RFLP (restriction fragment length polymorphism) data. RFLP, also known as DNA fingerprinting, provides a relatively inexpensive way to measure the similarity of isolates. It works by cutting DNA with restriction enzymes and then separating the resulting DNA fragments by running them through an electric field on a gel. Different sized fragments travel at different speeds down the gel producing a pattern of bands. The restriction enzymes are specific to certain short substrings of DNA, so if one sequence has insertions, deletions, or point mutations relative to another it may be cut in different places, producing its own distinctive spectrum of fragment sizes. Similar sequences have similar patterns of bands after being run on a gel. Gels can be aligned and scored according to the strength of bands of a particular molecular weight. For example, the scheme used by Schmid et al. [86] to compute dissimilarities for *C. albicans* based on the RFLP was

$$d(a,b) = 1 - \frac{\sum_{i=1}^{B}(a_i + b_i - |a_i - b_i|)}{\sum_{i=1}^{B}(a_i + b_i)}$$

Where, $B$ is the number of bands, and $x_i \in \{0,1,2,3\}$ is the intensity of band $i$ in isolate $x$. To see why such dissimilarities do not necessarily obey the triangle inequality consider the following example, where $d(a,b) = d(b,c) = \frac{1}{3}$, and $d(a,c) = 1$. So $d(a,b) + d(b,c) < d(a,c)$.

| Isolate | a | b | c |
|---|---|---|---|
| Band 1 | 1 | 1 | 0 |
| Band 2 | 0 | 1 | 1 |

Unlike sequence alignments, the characters in RFLP are not strongly linked. Instead, the entire genomic material is digested[2] (cut by restriction enzymes). Also, the characters are not independent, an addition of a restriction site creates two bands of a smaller molecular weight and will remove a band of a larger molecular weight [96].

Trees were constructed for the example data sets to identify clusters. Because, the dissimilarity information for each data sets cannot be exactly represented by a tree (as the data is not additive), it may be misleading to choose model strains based purely on this presentation of the data. However, displaying the choices of model strains, for the different criteria, on these trees helps to visualise how well they represent the diversity amongst the isolates. All trees were all built using neighbor-joining [83]. Neighbor-joining was used as it is a polynomial time algorithm that takes distances (or dissimilarities) as input.

### 5.4.1 *Pseudomonas areuginosa*

*P. aeruginosa* is a bacteria, like the yeast *C. albicans*, it is an opportunistic pathogen. It is capable of infecting a wide range of tissue types in people with impaired immune systems. The source of the *P. aeruginosa* data used here is Al-Samarrai et al. (1999) [85]. It is derived from RFLP based on *Sal*1 digests of genomic DNA from 22 isolates.

The neighbor-joining tree for *P. aeruginosa* is shown in figure 5.4. The tree does not contain any groups of highly similar isolates, and has no long internal edges that divide the isolates into natural clusters. There are two pairs of identical

[2] Sometimes plasmid DNA is separated from nuclear DNA (Jan Schmid, personal communication)

isolates within the data, i.e. 20 unique isolates. It was considered important to do the analysis on all isolates, rather than just unique isolates, as the extra information about which isolates are common is useful in determining model strains.

The number of model strains was fixed at $k = 1$ and the three dissimilarity based criteria (D1, D2, D3) were applied. Criteria D1 and D2 both choose either the isolate IAI8 or IAI27 as best model strain - these two isolates have identical banding patterns in the RFLP, i.e. d(IAI8,IAI27) = 0. Criteria D3 chose P6764. These strains are indicated in figure 5.4 by a blue circle (IAI8/IAI27) and a pink circle (P6764).

The small size of the data set (22 isolates), made it feasible to do a complete comparison of the exact criteria, and the greedy approximations to them, for $k = 1, 2, 3, 4$. A comparison of the costs and the model strains selected is shown for DC1, DC2, and DC3 in table 5.1. For DSC a comparison of $r^*$, the maximum number of isolates reachable, is shown in table 5.2 for different values of $k$ and $T$.

With this data, the costs of the greedy approximations to DC1, DC2 and DC3 were at most 10% more than the costs of the exact solutions. The model strains chosen by the greedy version of DSC reached at worst three fewer isolates (out of 22) than the exact solutions, for example with $T = 0.3$ and $k = 3$, $r^* = 13$ and $r$ for the greedy solution was 10.

The actual model strains picked by the dominating set method for $k = 1$ and different values of $T$, are given in table 5.3. They overlap with the choices of model strain for criteria DC1, DC2 and DC3.

The quartet based criterion for choosing model strains was not used for this data set because it requires the isolates to be split into predefined clusters. In this case there were no obvious clusters.

### 5.4.2 *Helicobacter pylori*

*H. pylori* is a species of bacteria that lives in the human gut, it can be the cause

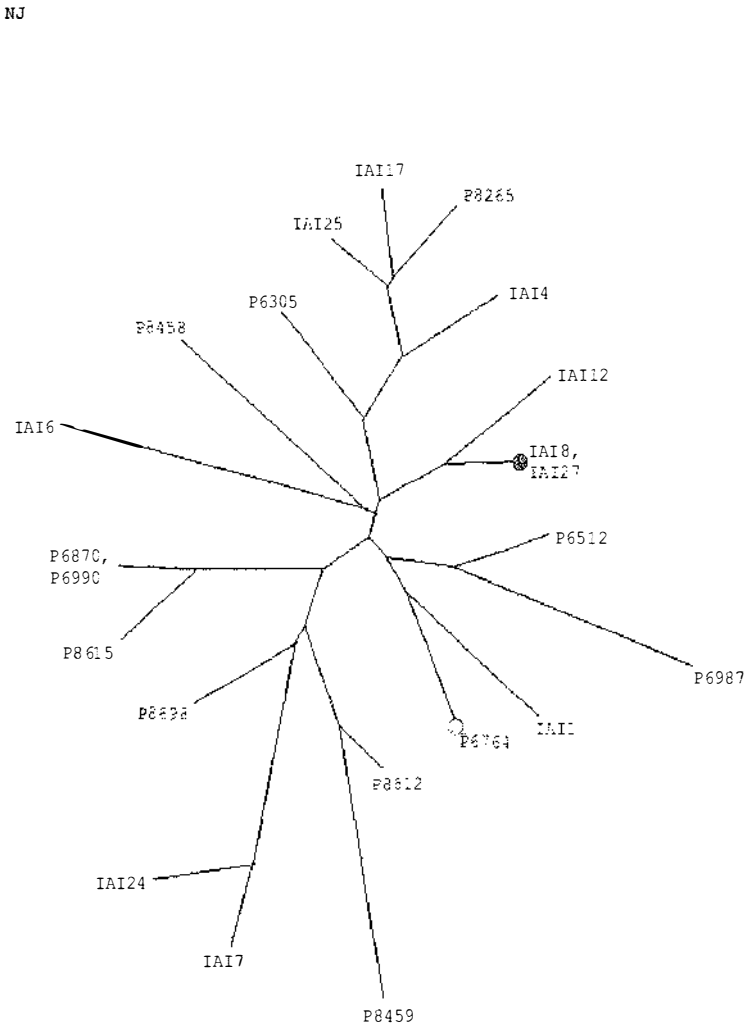


Figure 5.4: The neighbor-joining tree for 22 *P. aeruginosa* isolates, constructed using PAUP* [97]. The pink circle shows the minimum cost model strain ($k = 1$) for criteria D1 and D2. The blue circle marks the best model strain under D3.

DC1 (minimises the sum of dissimilarities)

| | $k=1$ | $k=2$ | $k=3$ | $k=4$ |
|---|---|---|---|---|
| Exact $C_k^*$ | 8.43 | 7.03 | 5.76 | 4.94 |
| M | IAI8 | IAI8, IAI25 | IAI8, IAI25, P8615 | IAI7, IAI8, IAI25, P6870 |
| Greedy $C_k^*$ | 8.43 | 7.03 | 5.76 | 5.02 |
| M | IAI8 | IAI8, IAI25 | IAI8, IAI25, P8615 | IAI7, IAI8, IAI25, P8615 |

DC2 (minimises the sum of squared dissimilarities)

| | $k=1$ | $k=2$ | $k=3$ | $k=4$ |
|---|---|---|---|---|
| Exact $C_k^*$ | 3.71 | 2.80 | 2.05 | 1.66 |
| M | IAI8 | IAI1, P8698 | IAI25, P6512, P8698 | IAI7, IAI8, IAI25, P6512 |
| Greedy $C_k^*$ | 3.71 | 2.92 | 2.12 | 1.74 |
| M | IAI8 | IAI8, IAI25 | IAI8, IAI25, P8615 | IAI7, IAI8, IAI25, P8615 |

DC3 (minimises the maximum dissimilarity)

| | $k=1$ | $k=2$ | $k=3$ | $k=4$ |
|---|---|---|---|---|
| Exact $C_k^*$ | 0.57 | 0.47 | 0.44 | 0.39 |
| M | P6764 | P6764, P8698 | P6305, P6512, P8698 | IAI4, IAI6, P6512, P8698 |
| Greedy $C_k^*$ | 0.57 | 0.47 | 0.46 | 0.43 |
| M | P6764 | P6764, P8698 | IAI6, P6764, P8698 | IAI6, P6512, P6764, P8698 |

Table 5.1: Exact and Greedy choices of model strains for *P. aeruginosa* using criteria DC1, DC2, and DC3. With DC1 the greedy algorithm gives identical answers to the exact algorithm for $k = 1, 2, 3$.

| | $k=1$ | $k=2$ | $k=3$ | $k=4$ |
|---|---|---|---|---|
| $T=0.1$ | 2 (2) | 4 (2) | 5 (4) | 6 (4) |
| $T=0.2$ | 4 (4) | 7 (7) | 9 (7) | 11 (9) |
| $T=0.3$ | 5 (5) | 10 (10) | 13 (10) | 16 (14) |
| $T=0.4$ | 11 (11) | 18 (15) | 21 (20) | 22 (21) |
| $T=0.5$ | 19 (19) | 22 (22) | 22 (22) | 22 (22) |
| $T=0.6$ | 22 (22) | 22 (22) | 22 (22) | 22 (22) |

Table 5.2: The exact and greedy values of $r^*$ (DSC) for different number of model strains $k$, and threshold values $T$ for *P. aeruginosa*. Greedy values are shown in brackets beside the exact values.

| $T$ | Model Strain/s | $r^*$ |
|---|---|---|
| 0.17–0.27 | IAI25 | 4 |
| 0.28–0.29 | IAI8 or IAI27 | 5 |
| 0.30–0.33 | IAI8, IAI25 or IAI27 | 5 |
| 0.34–0.37 | IAI1 or P6512 | 10 |
| 0.38 | P6512 | 11 |
| 0.39–0.41 | P6512 or P8698 | 11 |
| 0.42–0.45 | P6305 | 13-16 |
| 0.46–0.57 | IAI8 or IAI27 | 18-21 |
| 0.58 | P6764 | 22 |

Table 5.3: The best model strains for *P. aeruginosa*, $k=1$, according to DSC, for different values of $T$. $T$ took values between 0.17 and 0.58, in steps of 0.01. 0.17 is the smallest non-zero value in the dissimilarity matrix, $T=0.58$ is the first value for which all the isolates $i \in X$ have $D(i,m) < T$. Note that the model strain chosen for this value of $T$ is, naturally, the same as the one chosen by DC3. Where there were ties, all optimal model strains are shown in the table.

of peptic ulcers [2]. The source of the *H. pylori* data set used was Campbell et al. (1997) [13], the data contains 91 isolates. The dissimilarities are calculated from RFLP of *Hae*III digests of PCR products amplified from the *ureA-ureB* region.

The neighbor-joining tree for *H. pylori* is shown in figure 5.5. The data contains several different groups of identical isolates which makes it hard to label in a readable fashion, to simplify, the figure labels have been removed from the tree.

The lower half of the tree looks less diverse than the upper half which contains many long edges. The longer edges typically end with individual isolates rather than groups of identical isolates, these individuals are poor choices for model strains.

The results of DC1, DC2 and DC3 are shown in table 5.4 for $k = 1, 2, 3, 4$. The costs $C_k$ for the greedy algorithms are not more than 15% greater than the costs of the exact algorithms. The worst ratio occurs for DC3, $k = 2$, where the greedy $C_k$ is 15% more than $C_k^*$. The results for DSC are shown in table 5.5. Only the greedy version of the algorithm was used.

The bold line cutting the *H. pylori* tree in figure 5.5 defines the two clusters used for the quartet based criterion, the best model strain for each cluster is indicated on the tree by a purple triangle. Also shown on the tree are the model strains chosen by the three dissimilarity based criteria, in each case $k = 4$. The model strains for DC1, DC2, and DC3 are marked by blue, yellow and green circles respectively. Only 3 model strains are shown for DC3, as there was no improvement in the cost function from $k = 3$ to $k = 4$. From table 5.5 two sets of four model strains for the greedy version of DSC are shown, for the first set (shown by blue diamonds) the threshold was $T = 0.6$, and for the second set (shown by red diamonds) it was $T = 0.1$.

DC1, DC2, DC3 and DSC pick overlapping sets of model strains that appear to represent the major clusters in the tree. The quartet based criterion selects model strains that are at the end of long edges on the tree. This may be due to the problem illustrated in figure 5.2.

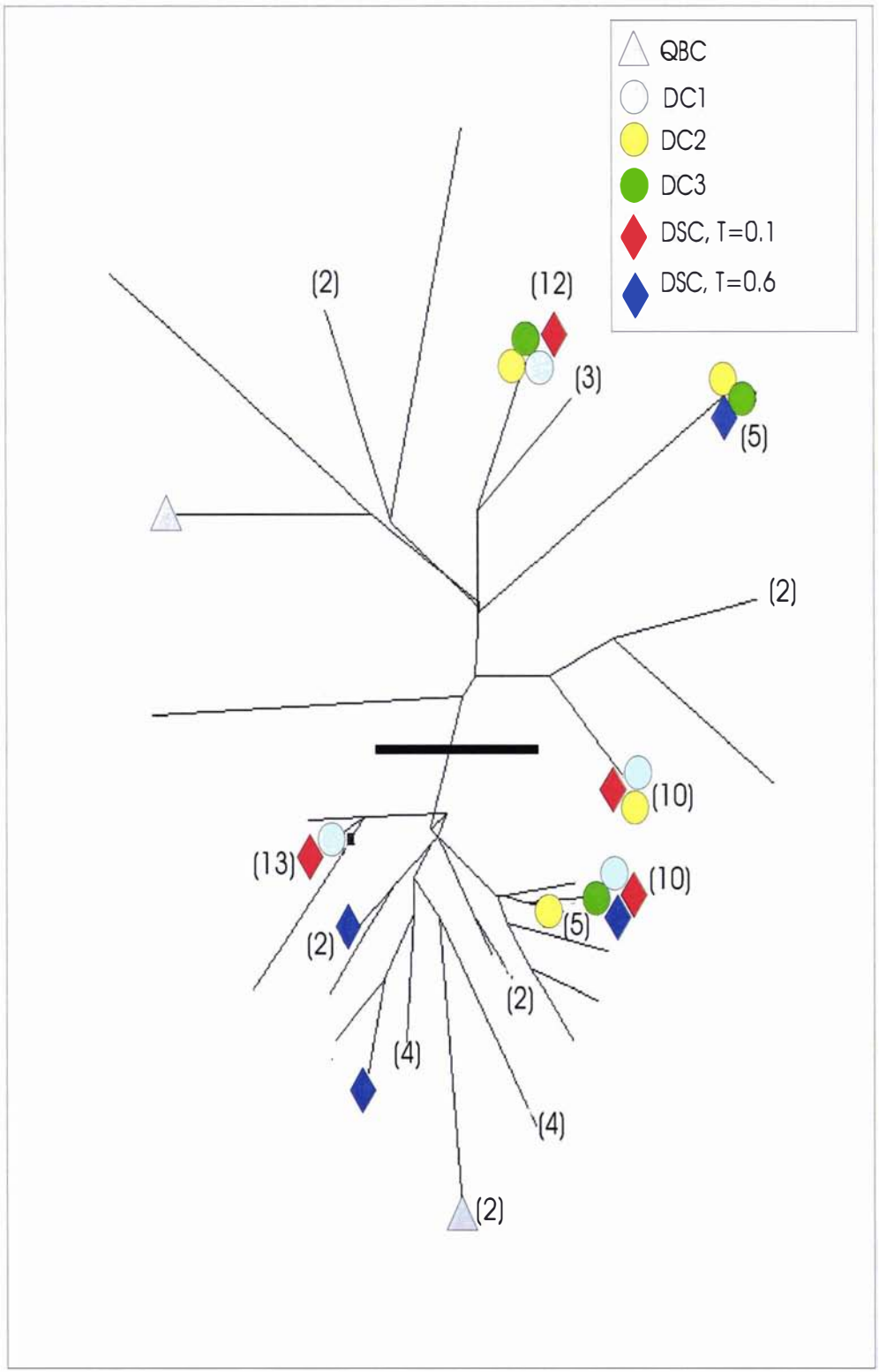


Figure 5.5: The neighbor-joining tree for 91 *H. pylori* isolates, constructed using PAUP* [97]. The labels have been removed to reduce clutter. The numbers in brackets indicate the number of identical isolates at each external node, when this is greater than one. The key in the top right indicates which symbols mark the model strains chosen by the different methods. The bold line cutting the tree divides the isolates into the two clusters that were used for QBC.

DC1 (minimises the sum of dissimilarities)

| | $k=1$ | $k=2$ | $k=3$ | $k=4$ |
|---|---|---|---|---|
| Exact $C_k^*$ | 36.56 | 27.08 | 21.53 | 17.88 |
| M | 3 | 4, 24 | 4, 14, 24 | 3, 4, 5, 14 |
| Greedy $C_k^*$ | 36.56 | 28.14 | 22.28 | 17.88 |
| M | 3 | 3, 4 | 3, 4, 5 | 3, 4, 5, 14 |

DC2 (minimises the sum of squared dissimilarities)

| | $k=1$ | $k=2$ | $k=3$ | $k=4$ |
|---|---|---|---|---|
| Exact $C_k^*$ | 19.38 | 12.11 | 9.03 | 6.97 |
| M | 3 | 4, 24 | 4, 14, 24 | 2, 4, 14, 24 |
| Greedy $C_k^*$ | 19.38 | 13.22 | 10.14 | 8.00 |
| M | 3 | 3, 4 | 3, 4, 14 | 2, 3, 4, 14 |

DC3 (minimises the maximum dissimilarity)

| | $k=1$ | $k=2$ | $k=3$ | $k=4$ |
|---|---|---|---|---|
| Exact $C_k^*$ | 1 | 0.67 | 0.60 | 0.60 |
| M | 1 | 4, 5 | 2, 4, 5 | 1, 2, 4, 5 |
| Greedy $C_k^*$ | 1 | 0.75 | 0.67 | 0.60 |
| M | 1 | 1, 4 | 1, 4, 5 | 1, 2, 4, 5 |

Table 5.4: Exact and Greedy choices of model strains for *H. pylori* using criteria DC1, DC2, and DC3.

| | $k=1$ | $k=2$ | $k=3$ | $k=4$ |
|---|---|---|---|---|
| $T=0.1$ | 13 | 25 | 35 | 45 |
| $T=0.2$ | 16 | 30 | 42 | 52 |
| $T=0.3$ | 35 | 47 | 57 | 65 |
| $T=0.4$ | 44 | 59 | 69 | 74 |
| $T=0.5$ | 60 | 74 | 79 | 82 |
| $T=0.6$ | 77 | 81 | 86 | 87 |
| $T=0.7$ | 80 | 90 | 91 | 91 |
| $T=0.8$ | 90 | 91 | 91 | 91 |
| $T=0.9$ | 90 | 91 | 91 | 91 |
| $T=1.0$ | 91 | 91 | 91 | 91 |

Table 5.5: The values of $r^*$ for different number of model strains $k$, and threshold values $T$ for *H. pylori*, as chosen by the greedy algorithm.

### 5.4.3 *Candida albicans*

*C. albicans*, a yeast, is an opportunistic pathogen best known for causing thrush. It is a common inhabitant of most humans, and usually causes no ill effects, however, in people with reduced immune systems, for example chemotherapy patients and AIDS sufferers, it can cause serious disease [51, 70]. It is thought to be mainly asexual, although recently this has been the matter of some debate [38, 76, 101].

The data contains RFLP banding patterns for 266 isolates. A neighbor-joining tree [83] of the data is shown in figure 5.6. The cluster identified by Schmid et al. [86] is indicated by the bold line.

The results of DC1, DC2 and DC3 are shown in table 5.6 for $k = 1, 2, 3, 4$. For $k = 1, 2, 3$ the greedy solutions do not cost much more than the exact solutions. Computing the exact solutions for $k = 4$ was too time consuming, extrapolating from the times for smaller $k$, suggests it would have taken approximately 300 hours per criterion. The results for DSC are shown in table 5.7. Only the greedy version of the algorithm was used.

The best model strain for each cluster according to QBC is indicated on the tree by a purple triangle. Also shown on the tree are the model strains chosen by the three distance based criteria, in each case $k = 3$. From table 5.7 two sets of

three model strains for the greedy dominating set based method are shown, for the first the threshold $T = 0.2$, and for the second set $T = 0.4$.

## 5.5 Discussion

To my knowledge, (micro)biologists do not employ any standard quantitative approach when selecting model strains for research. This chapter presented a range of criteria that could be useful to biologists seeking to identify representative model strains. Five different criteria were suggested and tested on three microbiological data sets of medical interest.

Of the criteria described, DC1, DC2, and DC3 were the simplest conceptually. DC3, which seeks to minimise the maximum distance of any isolate from its closest model strain, will prefer to choose model strains that are the least distance to outlying isolates, rather than choosing model strains from major clusters (subspecies). This may not be a desirable property for a model strain selection criterion.

The quartet based criterion, QBC, was not able to distinguish which members of a cluster were closest to the other members of the cluster. In the *H. pylori* example in particular it chose strains which were far from the complement of the cluster, these strains were not good representatives of their own clusters.

The modification of dominating set, DSC, gave more flexibility than the other criteria, with the specification of a threshold value $T$, this set a limit on how dissimilar a model strain could be from an isolate and still be considered representative. High threshold settings gave results similar to the criterion DC3, while lower values of $T$ gave results more similar to the criteria DC1 and DC2.

The three dissimilarity based criteria and the modification of dominating set are all computationally inefficient in their exact forms. This is due to the large number of sets of possible model strains. Greedy approximations to these criteria were developed, and performed well on the examples.

I close with a caveat, we have assumed that it is a good idea to chose as model

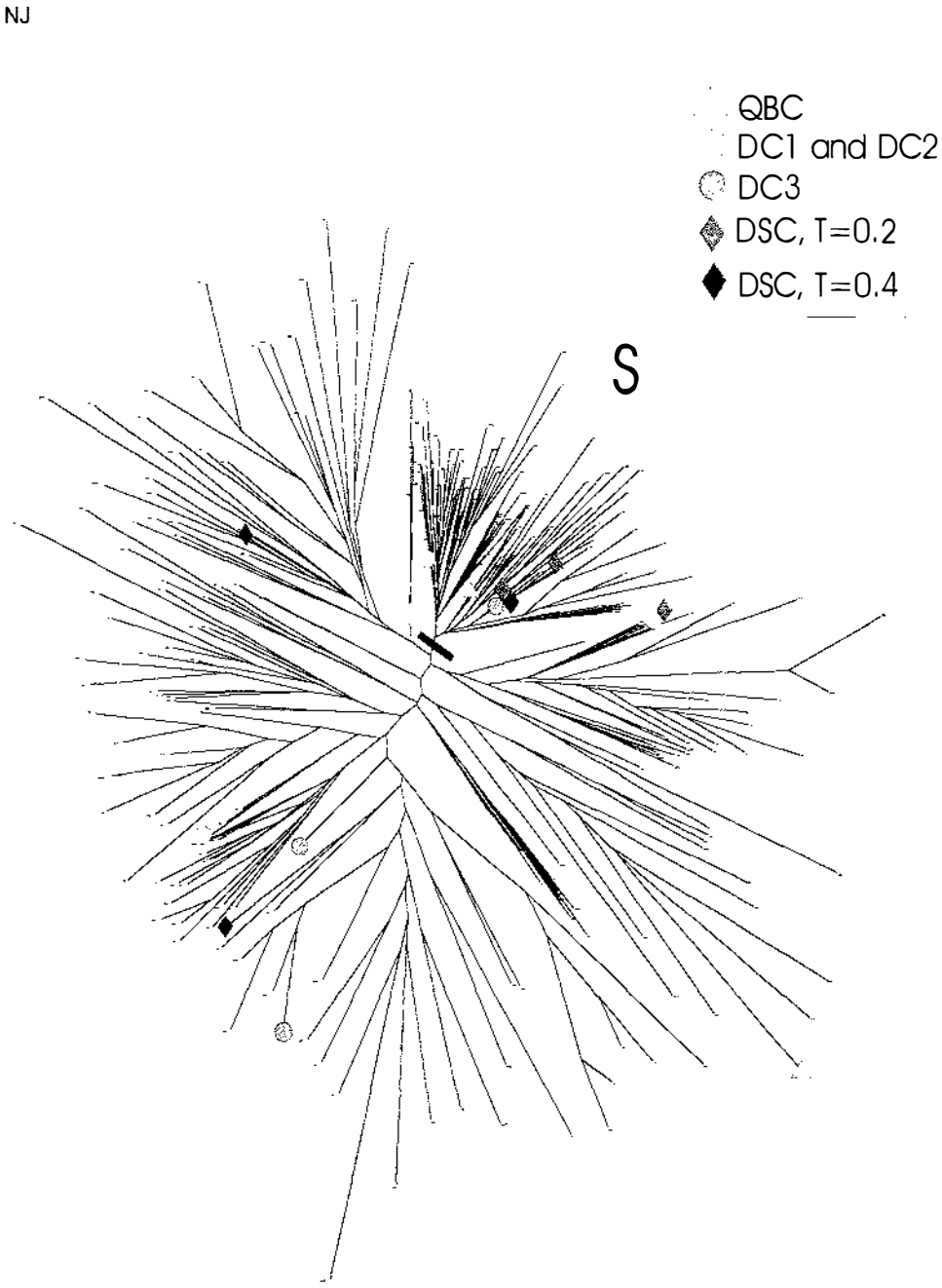


Figure 5.6: The neighbor-joining tree for 266 *C. albicans* isolates, constructed using PAUP* [97]. The key in the top right indicates the which symbols mark the model strains chosen by the different methods. The bold line cutting the tree indicates the cluster, marked $S$, defined by Schmid et al. [86]. It also divides the isolates into the two clusters that were used by QBC.

DC1 (minimises the sum of dissimilarities)

| | $k=1$ | $k=2$ | $k=3$ | $k=4$ |
|---|---|---|---|---|
| Exact $C_k^*$ | 65.02 | 57.42 | 54.45 | - |
| M | CH14 | W26, W55 | W26, W55, FJ27 | N.A. |
| Greedy $C_k^*$ | 65.02 | 58.72 | 55.34 | 53.11 |
| M | CH14 | CH14, Au35 | CH14, Au35, RIHO13 | CH14, Au35, RIHO13, Au36 |

DC2 (minimises the sum of squared dissimilarities)

| | $k=1$ | $k=2$ | $k=3$ | $k=4$ |
|---|---|---|---|---|
| Exact $C_k^*$ | 18.30 | 15.04 | 13.43 | - |
| M | CH14 | W26, W55 | W26, W55 FJ27 | N.A. |
| Greedy $C_k^*$ | 18.30 | 15.15 | 13.82 | 12.81 |
| M | CH14 | CH14, W79 | CH14, W79, CH35 | CH14, W79, CH35, FJ27 |

DC3 (minimises the maximum dissimilarity)

| | $k=1$ | $k=2$ | $k=3$ | $k=4$ |
|---|---|---|---|---|
| Exact $C_k^*$ | 0.48 | 0.41 | 0.40 | - |
| M | HUN122 | CH14, Au36 | CH14, W37, YsU63 | N.A. |
| Greedy $C_k^*$ | 0.48 | 0.45 | 0.42 | 0.40 |
| M | HUN122 | HUN122, W37 | HUN122, W37, Au152 | HUN122, W37, Au152, OTG3 |

Table 5.6: Exact and Greedy choices of model strains for *C. albicans* using criteria DC1, DC2, and DC3. For $k=4$ it took too long to compute the best model strains by exact search.

| | $k=1$ | $k=2$ | $k=3$ | $k=4$ |
|---|---|---|---|---|
| $T=0.1$ | 29 | 39 | 45 | 51 |
| $T=0.2$ | 96 | 131 | 144 | 153 |
| $T=0.3$ | 188 | 211 | 226 | 234 |
| $T=0.4$ | 252 | 262 | 265 | 266 |
| $T=0.5$ | 266 | 266 | 266 | 266 |

Table 5.7: The values of $r^*$ for different number of model strains $k$, and threshold values $T$ for *C. albicans*, as chosen by the greedy algorithm.

strains isolates that are somehow representative of their species or sub-species. However, in trying to avoid "oddball" strains we may not learn about the factors that make these strains different. These factors may include their reproductive strategy, or in the case of pathogens their virulence. One direction for future work, that could avoid this problem, would be to weight the isolates according to "interest", e.g. virulence.

# Bibliography

[1] A.C.Wilson, R.L. Cann, S.M. Carr, J. George, E.B. Gyllensten, K.M. Helmbychowski, R.G. Higuchi, S.R. Palumbi, E.M. Prager, R.D. Sage, and M. Stoneking. Mitochondrial DNA and two perspectives on evolutionary genetics. *Biol. J. Linn. Soc.*, 26:375–400, 1985.

[2] R.A. Alm and 22 others. Genomic-sequence comparison of two unrelated isolates of the human gastric pathogen *Helicobacter pylori* . *Nature*, 397:176 180, 1999.

[3] H.-J. Bandelt. Phylogenetic networks. *Verhandl. Naturwiss Vereins Hamburg*, 34:51–71, 1994.

[4] H.-J. Bandelt and A.W.M. Dress. A canonical decomposition theory for metrics on a finite set. *Adv. Math.*, 92:47–105, 1990.

[5] H.-J. Bandelt and A.W.M. Dress. Split decomposition: A new and useful approach to phylogenetic analysis of distance data. *Mol. Phyl. Evol.*, 1:242 252, 1992.

[6] H.-J. Bandelt, P. Forster, and A. Rohl. Median-joining networks for inferring intraspecific phylogenies. *Mol. Biol. Evol.*, 16:37–48, 1999.

[7] H.-J. Bandelt, P. Forster, B.C. Sykes, and M.B. Richards. Mitochondrial portraits of human populations using median networks. *Genetics*, 141:743–753, 1995.

[8] H.-J. Bandelt, V. Macaulay, and M. Richards. Median networks:speedy construction and greedy reduction, one simulation, and two case studies from human mtDNA. *Mol. Phyl. Evol.*, 16:8–28, 2000.

[9] C.W. Birky. Uniparental inheritance of mitochondrial and chloroplast genes: mechanisms and evolution. *Proc. Natl. Acad. Sci.*, 92:11331–11338, 1995.

[10] P.L. Bollyky, A. Rambaut, P.H. Harvey, and E.C. Holmes. Recombination between sequences of Hepatitis B virus from different genotypes. *J. Mol. Evol.*, 42:97–102, 1996.

[11] W.J. Bruno and A.L. Halpern. Topological bias and inconsistency of maximum likelihood using wrong models. *Mol. Biol. Evol.*, 16:564–566, 1999.

[12] P. Buneman. The recovery of trees from measures of dissimilarity. In *Mathematics in the Archaeological and Historical Sciences*, pages 387–395. Edinburgh University Press, 1971.

[13] S. Campbell, A. Fraser, B. Holliss, J. Schmid, and P. W. O'Toole. Evidence for ethnic tropism of *Helicobacter pylori* . *Infect. Immun.*, 65:3708–3712, 1997.

[14] M.A. Charleston. *Factors affecting the performance of phylogenetic methods.* PhD thesis, Massey University, 1994.

[15] M.A. Charleston, M.D. Hendy, and D. Penny. The effects of sequence length, tree topology, and number of taxa on the performance of phylogenetic methods. *J. Comp. Bio.*, 1:133–151, 1994.

[16] B. Chor, M.D. Hendy, B.R. Holland, and D. Penny. Multiple maxima of likelihood in phylogenetic trees: An analytic approach. *Mol. Biol. Evol.*, 17:1529–1541, 2000.

[17] A. Cooper and D. Penny. Mass survival of birds across the Cretaceous-Tertiary boundary: Molecular evidence. *Science*, 275:1109–1113, 1997.

[18] Z. Debyser, E.v. Wijngaerden, K.v. Laethem, K. Beuselinck, M. Reynders, E.D. Clerq, J. Desmyter, and A.M. Vandame. Failure to quantify viral load with two of the three commercial methods in a pregnant woman harboring a HIV type 1 subtype G strain. *AIDS Research and Human Retroviruses*, 14:453–459, 1998.

[19] A. Dress. Statistische geometrie von konfigurationen und deren evolution in sequenz-räumen - definitionen und probleme. ein programmvorschlag. In *Die Bedentung der von Berlin aus gehenden Mathematik in Vergangenheit und Gegenwart.* Kolloquium-Verlag, Berlin., 1988.

[20] A. Dress, B. Holland, K.T. Huber, J.H. Koolen, V. Moulton, and J. Weyer-Menkoff. $\delta$ additive and $\delta$ ultra-additive maps, Gromov's trees, and the Farris transform. submitted to Discrete Applied Math., 2001.

[21] A. Dress, D. Huson, and V. Moulton. Analyzing and visualizing sequence and distance data using SplitsTree. *Disc. Appl. Math.*, 71:95–109, 1996.

[22] A. Drummond, G.K. Nicholls, A.G. Rodrigo, and W. Solomon. Estimating mutation rate, population history, substitution model and genealogy simultaneously from temporally spaced sequence data. *Unpublished*, 2001.

[23] A. Drummond and A.G. Rodrigo. Reconstructing genealogies of serial samples under the assumption of a molecular clock using serial sample UPGMA (sUPGMA). *Mol. Biol. Evol.*, 17:1807–1815, 2000.

[24] E.S. Edgington. *Randomization Tests.* Marcel Dekker, Inc., 3rd edition, 1995.

[25] B. Efron. *The Jackknife, the Bootstrap and other resampling plans.* Dept. of Statistics Stanford University, 1982.

[26] B. Efron, E. Halloran, and S. Holmes. Bootstrap confidence levels for phylogenetic trees. *Proc. Natl. Acad. Sci.*, 93:13429–13434, 1996.

[27] B. Efron and R.J. Tibshirani. *An introduction to the bootstrap.* Chapman and Hall, 1993.

[28] M. Eigen and R. Winkler-Oswatitsch. Statistical geometry on sequence space. *Methods in Enzymology*, 183:505–530, 1990.

[29] M. Eigen, R. Winkler-Oswatitsch, and A. Dress. Statistical geometry in sequence space: a method of quantitative sequence analysis. *Proc. Natl. Acad. Sci.*, 85:5913–5917, 1988.

[30] B.S. Everitt. *Cluster Analysis.* Edward Arnold, 3rd edition, 1993.

[31] J.S. Farris. A probability model for inferring evolutionary trees. *Syst. Zool.*, 22:250–256, 1973.

[32] J. Felsenstein. Cases in which parsimony or compatibility methods will be positively misleading. *Syst. Zool.*, 27:401–410, 1978.

[33] J. Felsenstein. Evolutionary trees from DNA sequences: A maximum likelihood approach. *J. Mol. Evol.*, 17:368–376, 1981.

[34] J. Felsenstein. *PHYLIP (Phylogeny Inference Package) version 3.5c.* Department of Genetics, University of Washington, Seattle, 1993. Distributed by the author.

[35] W.M. Fitch. Toward defining the course of evolution: Minimum change for a specific tree topology. *Syst. Zool.*, 20:406–416, 1971.

[36] M.R. Garey and D.S. Johnson. *Computers and intractability: A guide to the theory of NP-completeness.* W.H. Freeman and Company, 1979.

[37] A.D. Gordon. *Classification: methods for exploratory analysis of multivariate data.* Chapman and Hall, 1981.

[38] Y. Graser, M. Volovsek, J. Arrington, G. Schonian, W. Presber, T.G. Mitchell, and R. Vilgalys. Molecular markers reveal that population structure of the human pathogen *Candida albicans* exhibits both clonality and recombination. *Proc. Nat. Acad. Sci.*, 93:12473–12477, 1996.

[39] N.C. Grassly and E.C. Holmes. A likelihood method for the detection of selection and recombination using nucleotide sequences. *Mol. Biol. Evol.*, 14:239–247, 1997.

[40] M. Gromov. Hyperbolic groups. In *Essays in group theory.*, volume 8. Springer-Verlag, 1988.

[41] M. Hasegawa, H. Kishino, and T. Yano. Dating of the human-ape splitting by a molecular clock of mitochondrial DNA. *J. Mol. Evol.*, 21:160–174, 1985.

[42] J.C. Heine and T.W. Speir. Ornithogenic soils of the Cape Bird Adélie penguin rookeries, Antarctica. *Geology*, 22:23–26, 1994.

[43] M.D. Hendy and D. Penny. A framework for the quantitative study of evolutionary trees. *Syst. Zool.*, 38:297–309, 1989.

[44] M.D. Hendy and D. Penny. Spectral analysis of phylogenetic data. *J. Class.*, 10:5–24, 1993.

[45] M.D. Hendy, D. Penny, and M.A. Steel. A discrete fourier analysis for evolutionary trees. *Proc. Natl. Acad. Sci.*, 91:3339–3343, 1994.

[46] D.M. Hillis, J.P. Huelsenbeck, and C.W. Cunningham. Application and accuracy of molecular phylogenies. *Science*, 264:671–677, 1994.

[47] E.C. Holmes, M. Worobey, and A. Rambaut. Phylogenetic evidence for recombination in Dengue virus. *Mol. Biol. Evol.*, 16:405–409, 1999.

[48] M. Hoss, P. Jaruga, T. Zastawny, M. Dizdaroglu, and S. Pääbo. DNA damage and DNA sequence retrieval from ancient tissues. *Nucleic Acids Res.*, 24:1304–1307, 1996.

[49] D. Huson. *SplitsTree version 2.4*, 1997. Available from **ftp://ftp.uni-bielefeld.de/pub/math/splits/splitstree2**.

[50] D. Huson. SplitsTree: a program for analyzing and visualizing evolutionary data. *Bioinformatics*, 14:68–73, 1998.

[51] M.M. Jensen and D.N. Wright. *Introduction to microbiology for the health sciences.* Prentice Hall, 3rd edition, 1993.

[52] J. Kim. Improving the accuracy of phylogenetic estimation by combining different methods. *Syst. Biol.*, 42:331-340, 1993.

[53] J. Kim. Slicing hyperdimensional oranges: the geometry of phylogenetic estimation. *Mol. Phyl. Evol.*, 17:58–75, 2000.

[54] M. Kimura. Estimation of evolutionary sequences between homologous nucleotide sequences. *Proc. Natl. Acad. Sci.*, 78:454–458, 1981.

[55] J.F.C. Kingman. The coalescent. *Stochastic Processes and their Applications*, 13:235–248, 1982.

[56] J.F.C. Kingman. On the genealogy of large populations. *J. Appl. Prob.*, 19A:27–43, 1982.

[57] S. Kumar, K. Tamura, and M. Nei. *MEGA: Molecular Evolutionary Genetics Analysis, version 1.01.* The Pennsylvania State University, University Park, PA 16802, 1993.

[58] D.M. Lambert, P.A. Ritchie, C.D. Millar, B. Holland, A. Drummond, and C. Baroni. Evolution in action: Ancient DNA from frozen Adélie penguin bones in Antarctica. Submitted to Nature, 2001.

[59] T. Lindahl. Instability and decay of the primary structure of DNA. *Nature*, 362:709-715, 1993.

[60] P.J. Lockhart and S.A. Cameron. Trees for bees. *Trends in Eco. and Evol.*, 16:84–88, 2001.

[61] J. Lyons-Weiler, G.A. Hoelzer, and R.J. Tausch. Relative Apparent Synapomorphy Analysis (RASA) I: The statistical measurement of phylogenetic signal. *Mol. Biol. Evol.*, 13:749-757, 1996.

[62] J. Lyons-Weiler, G.A. Hoelzer, and R.J. Tausch. Optimal outgroup analysis. *Biol. J. Linn. Soc.*, 64:493–511, 1998.

[63] S. Marchant and P.J. Higgins. *Handbook of Australia, New Zealand and Antarctic Birds.* Oxford University Press, Melbourne, 1990.

[64] G. McGuire, F. Wright, and M. J. Prentice. A graphical method for detecting recombination in phylogenetic data sets. *Mol. Biol. Evol.*, 14:1125–1131, 1997.

[65] S. Meyer, G. Weiss, and A.v. Haeseler. Patterns of nucleotide substitution and rate heterogeneity in the hypervariable regions I and II of human mtDNA. *Genetics*, 152:1103–1110, 1999.

[66] M.C. Milinkovitch. Finding optimal ingroup topologies and convexities when the choice of outgroups is not obvious. *Mol. Phyl Evol.*, 9:348-357, 1997.

[67] D.P. Mindell, M.D. Sorenson, D.E. Dimcheff, M. Hasegawa, J.C. Ast, and T. Yuri. Interordinal relationships of birds and other reptiles based on whole mitochondrial genomes. *Syst. Biol.*, 48:138–152, 1999.

[68] K. Nieselt-Struwe and A. von Haeseler. Quartet-mapping, a generalization of the likelihood mapping procedure. *Mol. Biol. Evol.*, 18:1204–1219, 2001.

[69] E.W. Noreen. *Computer intensive methods for testing hypotheses: An introduction.* John Wiley and Sons, 1989.

[70] F.C. Odds. Candida *and Candidosis.* Bailliere Tindall, 2nd edition, 1988.

[71] S. Pääbo. Ancient DNA:Extraction, characterization, molecular cloning, and enzymatic amplification. *Proc. Natl. Acad. Sci.*, 86:1939–1943, 1989.

[72] D. Penny and M. Hasegawa. The platypus put in its place. *Nature*, 387:549–550, 1997.

[73] D. Penny, M.D. Hendy, P.J. Lockhart, and M.A. Steel. Corrected parsimony, minimum evolution, and Hadamard conjugations. *Syst. Biol.*, 45:596–606, 1996.

[74] D. Penny, M.D. Hendy, and M.A. Steel. Testing the theory of descent. In *Phylogenetic analysis of DNA sequences*, pages 155–183. Oxford University Press, Inc., 1991.

[75] D. Posada and K.A. Crandall. Simple (wrong) models for complex trees: a case from retroviridae. *Mol. Biol. Evol.*, 18:271–275, 2001.

[76] C. Pujol, J. Reynes, F. Renaud, M. Raymond, M. Tibayrenc, F.J. Ayala, F. Janbon, M. Mallie, and J. Bastide. The yeast *Candida albicans* has a clonal mode of reproduction in a population of infected human immunodeficiency virus-positive patients. *Proc. Nat. Acad. Sci.*, 90:9456–9459, 1993.

[77] Y. Qui, J. Lee, F. Bernasconi-Quadroni, D.E. Soltis, P.S. Soltis, M. Zanis, E.A. Zimmer, Z. Chen, V. Savolainen, and M.W. Chase. The earliest angiosperms: evidence from mitochondrial, plastid and nuclear genomes. *Nature*, 402:404–407, 1999.

[78] T.W. Quinn. The genetic legacy of Mother Goose - phylogeographic patterns of lesser snow goose *Chen caerulescens caerulescens* maternal lineages. *Mol. Ecol.*, 1:105–117, 1992.

[79] A. Rambaut. Estimating the rate of molecular evolution: incorporating non-contemporaneous sequences into maximum likelihood phylogenies. *Bioinformatics*, 16:395–399, 2000.

[80] A.E. Rambaut and N.C. Grassly. Treevolve version 1.3. Available from **http://evolve.zoo.ox.ac.uk**.

[81] A.E. Rambaut and N.C. Grassly. Seq-Gen: An application for the Monte Carlo simulation of DNA sequence evolution along phylogenetic trees. *Comput. Appl. Biosci.*, 1997.

[82] P. Ritchie. *The evolution of the mitochondrial DNA control region in the Adélie penguins of Antarctica.* PhD thesis, Massey University, New Zealand, 2001.

[83] N. Saitou and T. Imanishi. Relative efficiencies of the fitch-margoliash, maximum-parsimony, maximum-likelihood, minimum-evolution, and neighbor-joining methods of phylogenetic tree construction in obtaining the correct tree. *Mol. Biol. Evol.*, 6:514 -525, 1989.

[84] N. Saitou and M. Nei. The Neighbor-joining method: A new method for reconstructing phylogenetic trees. *Mol. Biol. Evol.*, 4:406–425, 1987.

[85] T. Al Samarrai, N. Zhang, I. Lamont, J. Kolbe, M. Wilsher, A. J. Morris, and J. Schmid. A fast, highly discriminating and inexpensive method for computer-assisted typing of *Pseudomonas aeruginosa.* NZMS meeting, Dunedin, New Zealand, 1999.

[86] J. Schmid, S. Herd, P.R. Hunter, R.D. Cannon, and M. Salleh. Evidence for a general-purpose genotype in *Candida albicans* , highly prevalent in multiple geographical regions, patient types and types of infection. *Microbiology*, 145:2405–2413, 1999.

[87] S. Siguroardóttir, A. Helgason, J.R. Gulcher, K. Stefansson, and P. Donnelly. The mutation rate in the human mtDNA control region. *Am. J. Hum. Genet.*, 66:1599–1609, 2000.

[88] A.B. Smith. Rooting molecular trees: problems and strategies. *Biol. J. Linn. Soc.*, 51:279–292, 1994.

[89] R.R. Sokal and C.D. Michener. A statistical method for evaluating systematic relationships. *Univ. Kansas Sci. Bul.*, 38:1409–1438, 1958.

[90] P.S. Soltis, D.E. Soltis, and M.W. Chase. Angiosperm phylogeny inferred from multiple genes as a tool for comparative biology. *Nature*, 402:402–404, 1999.

[91] J. Sourdis and C. Krimbas. Accuracy of phylogenetic trees estimated from DNA sequence data. *Mol. Biol. Evol.*, 4:159–166, 1987.

[92] J. Sourdis and M. Nei. Relative efficiencies of the maximum parsimony and distance-matrix methods in obtaining the correct phylogenetic tree. *Mol. Biol. Evol.*, 5:298–311, 1988.

[93] M.A. Steel, M.D. Hendy, and D. Penny. Parsimony can be consistent! *Syst. Biol.*, 42:581–587, 1993.

[94] C.K. Stover and 30 others. Complete genome sequence of *Pseudomonas aeruginosa* PA01, an opportunistic pathogen. *Nature*, 406:959–964, 2000.

[95] K. Strimmer and A. von Haesler. Likelihood-mapping: a simple method to visualize phylogenetic content of sequence alignment. *Proc. Natl. Acad. Sci. USA*, 94:6815–6819, 1997.

[96] D. Swofford, G. Olsen, P. Waddell, and D. Hillis. Phylogenetic inference. In *Molecular Systematics*, pages 407–514. Sinauer Associates, 1996. 2nd edition.

[97] D.L. Swofford. *PAUP* - Phylogenetic Analysis Using Parsimony (*and Other Methods)*, 1998. Version 4. Sinauer Associates, Sunderland, Mass.

[98] F. Tajima. Unbiased estimation of evolutionary distance between nucleotide sequences. *Mol. Biol. Evol.*, 10:677–688, 1993.

[99] K. Talaro and A. Talaro. *Foundations in microbiology.* Dubuque, IA, 1996.

[100] Y. Tateno, N. Takezaki, and M. Nei. Relative efficiencies of the maximum-likelihood, neighbour joining and maximum parsimony methods when substitution rates varies with site. *Mol. Biol. Evol.*, 11:261–277, 1994.

[101] M. Tibayrenc. Are *Candida albicans* natural populations subdivided? *Trends in Microbiology*, 5:253-257, 1997.

[102] J.-F. Tomb and 41 others. The complete genomic sequence of the gastric pathogen *Helicobacter pylori* . *Nature*, 388:539 547, 1997.

[103] P.J. Waddell, Y. Cao, J. Hauf, and M. Hasegawa. Using novel phylogenetic methods to evaluate mammalian mtDNA, including amino acid-invariant sites-logdet plus site stripping, to detect internal conflicts in the data, with special reference to the positions of hedgehog, armadillo, and elephant. *Syst. Biol.*, 48:31- 53, 1999.

[104] P.J. Waddell, D. Penny, M.D. Hendy, and G. Arnold. The sampling distributions and covariance matrix of phylogenetic spectra. *Mol. Biol. Evol*, 11:630–642, 1994.

[105] P.J. Waddell, D. Penny, and T. Moore. Hadamard conjugations and modelling sequence evolution with unequal rates across sites. *Mol. Phyl. Evol*, 8:33–50, 1997.

[106] E.E. Watson. *Threads from the past: A genetic study of African ethnic groups and human origins.* PhD thesis, Massey University, New Zealand, 1996.

[107] G.F. Weiller. Phylogenetic profiles: A graphical method for detecting genetic recombinations in homologous sequences. *Mol. Evol. Sys.*, 15:326–335, 1998.

[108] B.S. Weir. *Genetic Data Analysis II.* Sinauer Associates Inc., 2nd edition, 1996.

[109] T.D. Williams. *The Penguins: Spheniscidae.* Oxford University Press, Oxford, 1995.

[110] Z. Yang. How often do wrong models produce better phylogenies? *Mol. Biol. Evol.*, 14:105–108, 1997.