UNIVERSITY OF LONDON THESIS

Degree PhD Year 2007 Name of Author DAVID CHARLES DALE

COMPARISONS OF PARSIMONY AND LIKELIHOOD-BASED METHODS IN PHYLOGENETIC ANALYSIS

DAVID DALE

DEPARTMENT OF BIOLOGY AND COMPLEX

UNIVERSITY COLLEGE LONDON

2006

SUBMITTED TO THE UNIVERSITY OF LONDON

FOR THE DEGREE OF DOCTOR OF PHILOSOPHY

ProQuest LLC
789 East Eisenhower Parkway
P.O. Box 1346
Ann Arbor, MI 48106-1346

ABSTRACT

This work provides a defence of the claim that likelihood based methods provide a better framework for performing phylogenetic analyses on molecular sequences than do parsimony based methods under the conditions studied.

Novel work introduced in the thesis includes simulation studies that examine the performance of likelihood based and parsimony based methods at high evolutionary distances. At these distances, many changes accumulate at a single site causing a catastrophic collapse in the performance of the parsimony analysis. In contrast a well understood mathematical theory involving the use of Fisher's information measure describes the decline in performance of likelihood methods.

Further work compares the performance of likelihood based methods and parsimony methods under heterotachous conditions, i.e. conditions under which a single site will alter its rate of evolution relative to other sites. A recent claim that parsimony based analyses outperform likelihood is rebutted and a likelihood model is introduced and its performance analysed.

Finally likelihood based methods are defined in terms of rates. A method for turning these rates into a probability distribution describing the number of changes of interest across a phylogeny is described. This is then compared to the number of changes inferred under a parsimony analysis. When the true model is known, it is shown that the counts of changes inferred under a parsimony based analysis have a low probability of being correct. It is argued that this accounts for the poor performance of parsimony.

ACKNOWLEDGEMENTS

I would like to thank my supervisor Professor Ziheng Yang without whose guidance and patience this would not have been possible. I am also very grateful to Dr. Joe Bielawski, Dr. Paul Agapow and Dr. Richard Emes for their advice, encouragement and support, as well as their good humour.

for Rebecca

CONTENTS

LIST OF FIGURES

INTRODUCTION

Recently it has become clear that an understanding of the evolutionary history of biological sequences is a great aid to both small and large scale biological sequence analyses, (RIVERO *et al.* 2005). The popularity of comparative sequence analysis of this kind owes its existence to the confluence of two rapidly developing technologies. Advances in sequencing technology have provided an abundance of data to analyse, whilst at the same time advances in computing technology have provided the means to analyse it.

Methods for performing evolutionary analyses can be divided into two broad categories. In the first category are cladistic methods, based on a minimum change criterion (FITCH 1971). Evolution is regarded as proceeding by a series of discrete events that change the state of the biological sequence under study. A minimum-change analysis is then performed by minimising the number of those evolutionary events and attempting to draw inferences from their pattern. In the second category fall the likelihood methods (WHELAN *et al.* 2001). These aim to explicitly model evolution as a probabilistic process. The aim of the analysis is then to estimate the parameters that define that process. The work presented here provides a defence of the claim that, under the conditions of this study, the statistical methods provide a better framework for performing evolutionary analyses on molecular sequences.

In this work we address two of the biological problems to which both categories of methods have been applied. These are: (1) the problem of reconstructing evolutionary histories or phylogenetic trees and (2) Kimura's Neutral Theory of molecular evolution (KIMURA 1983). The first problem has as its aim the reconstruction of the relationship between all species, both extant and extinct. Cladistic methods approach this problem by

choosing the evolutionary tree for which the data can be explained by the minimum inferred number of changes. Likelihood methods approach the problem by defining an explicit probabilistic model of sequence evolution and choosing the tree for which the probability of observing the data is maximised. The second problem is a counterpoint to Darwin's theory of natural selection, and states that most of the changes observed at the molecular level are selectively neutral and are driven not by selection but by random drift. This theory has been incorporated into evolutionary biology as a null hypothesis that researchers interested in adaptive mutations seek to reject. Cladistic methods test the hypothesis by counting inferred substitutions and statistical methods test the hypothesis by modelling the processes involved explicitly. An overview of these biological problems and the different methods used in analysing them is provided in this work. Previous comparisons that have been made are given, including both simulation studies and attempts to interpret cladistic analyses in terms of probabilities.

The work introduced here falls into two parts. The first part, covered in Chapters 3 and 4, aims to discover how cladistic and statistical methods perform when the analysis is known to be difficult. In the first case, we examine the performance of statistical and cladistic methods at large evolutionary distances. At these distances, many changes accumulate at a single site. Cladistic methods, based on a minimum change criterion, show a collapse in performance under these conditions. In contrast, though likelihood methods also show a decline in performance, the decline follows a well understood mathematical theory. We describe that theory, which involves the use of Fisher's information measure and show that it more or less accurately describes the decline in performance of likelihood methods.

In another case of a difficult evolutionary scenario, we investigate how the two classes of methods perform at tree reconstruction when the evolution of the sequences is heterotachous. In these cases a subset of sites on a sequence alter their rate of evolution relative to other sites. A recent claim (KOLACZKOWSKI and THORNTON 2004) that cladistic analyses outperform traditional non-heterotachous likelihood under these conditions is rebutted. Additionally a novel likelihood model is introduced and its performance analysed.

In the second part of the work, covered in Chapter 5, we introduce a practical method for turning the rates of change defined in a statistical analysis into a distribution that assigns a probability to the number changes that have occurred on a phylogeny, given the observed data. Thus given some data, an evolutionary tree and a probabilistic model, one can calculate the probability that a certain number of a subset of changes of interest have occurred. As well as being of interest in its own right, this is of use when investigating the difference in performance between statistical and cladistic analyses. The number of changes inferred under a cladistic minimum change criterion can be calculated and, when the true model is known, can be compared to the probability of that number of changes occurring. It is shown that the counts of changes inferred under a parsimony based analysis have a low probability of being correct. It is argued that this accounts for the poor performance of parsimony.

CHAPTER 1. BACKGROUND TO THE PROBLEMS ADDRESSED

This thesis addresses the comparison of two inference methods used in evolutionary biology. In this chapter we provide the background to the biology that underlies this thesis and a description of the biological problems that the inference methods have been used to address. In this section we also explain the meaning of some of the biological terms that are used in the rest of the thesis. We take as a starting point, Darwin's theory of natural selection (DARWIN 1859) and the central dogma of biology (CRICK 1970), that DNA makes RNA makes the proteins that ultimately determine an organism's traits. DNA replication provides the mechanism for the heritability of those traits.

Variety is generated by altering the information encoded by the DNA by changing the pattern of its constituent bases, i.e. the pattern of As, Ts, Gs and Cs. This alteration can occur either by large scale chromosomal rearrangements called recombination events (KREUZER 2005), by small insertion and deletion events that add or remove small numbers of bases from the copied strand or by single point mutations that alter the base at the copied strand (MAKI 2002). This last type of mutation event is the best understood. It can occur because bases will occasionally flip into an alternative tautomeric form and form bonds in ways that they would not usuallly do, for example a cytosine in a tautomeric form will bind to an adenine (STRAZEWSKI 1988). Also alternative bindings are sometimes stable, for example a guanine - tyrosine binding is stable if the helix is slightly distorted (CAI and CONNOLLY 1997). Finally point mutations can be caused by environmental factors such as ionising radiation (GROSOVSKY *et al.* 1988), alkylating agents such as S-adenosyl methionine - a product of glycolysis (MACINTYRE *et al.* 2001) - and free oxygen radicals (MURATA-KAMIYA *et al.* 1995). Point mutations that turn a purine into a pyrimidine, or

vice-versa, are known as transversions whilst those that mutate a purine into a purine or a pyrimidine into another pyrimidine are known as transitions. Point mutations have been much more extensively studied than the other classes of mutations and hence the methods investigated in this thesis involve point mutations exclusively. However some attempts have been made to handle small-scale insertions and deletions, eg (FLEISSNER *et al.* 2005; REDELINGS and SUCHARD 2005).

1.1 The biological problems

This thesis compares the performance of two different methods used for inference about evolutionary biology, investigating how they perform in situations that are biologically plausible but methodologically taxing. Here we describe the general biological problems addressed.

1.1.1 Patterns of descent

Perhaps the most obvious problem, given the biological theory of descent, is that of how to reconstruct the pattern of that descent. It seems apparent from their shared morphology that dogs and wolves share a common ancestor that occurs more recently than the common ancestor of cows, dogs and wolves. Equally it seems apparent that cows, dogs and wolves share a more recent common ancestor than do cows, dogs, wolves and herring. This pattern of ancestry can be represented as a tree of life, a metaphor noticed by (DARWIN 1859). One of the projects of evolutionary biology is to construct the tree of life for all organisms (MADDISON 2004). For the purpose of this work these trees of evolutionary relationships are called phylogenies, a term first coined by (HAECKEL 1866). It is unclear

exactly that Haeckel meant a phylogeny to merely mean a diagrammatic tree of evolutionary relationships. It seems (DAYRAT 2003) that his phylogenies, by his definition histories of the development of forms, conveyed a notion of progress that is absent in Darwinian thought. Here the difference is ignored. Different methods for constructing phylogenies will be described in a later chapter, and comparisons between them will form one strand of the results section of this thesis.

1.1.2 Non-Darwinian Evolution

We have seen that in the orthodox synthesis described, mutations in the DNA provide the raw material on which natural selection works. However it was noted that "Natural selection is ... the editor of the genetic message. One thing the editor does not do is to remove changes which it is unable to perceive"(KING and JUKES 1969). This, in a nutshell, is the neutral theory (KIMURA 1983). This proposes that mutations acumulate on DNA molecules at an extremely rapid rate, in his original study (KIMURA 1968) he put it at about 1 mutation per 10 million base pairs per generation. His argument was that these mutations were happening at such a high rate that they had to be neutral or nearly neutral. In other words they were becoming fixed in natural populations without providing a selective advantage. It was further argued that neutral or nearly neutral (OHTA 2002) mutations were becoming fixed at a rate equal to their mutation rate. Hence it was argued that most molecular evolutionary change comes not from the Darwinian process of natural selection but from the random accumulation and fixation of mutations that have no effect on the probability of DNA replication and hence are invisible to selection. The theory was met with hostility from those who believed selection played a dominant role in evolution (CLARKE 1970; RICHMOND 1970). They had some reason on their side; it has since been

argued that silent synonymous mutations can offer selective advantage in bacteria (LYNN *et al.* 2002).

Kimura's theory is incorporated into evolutionary biology as a null hypothesis which researchers may seek to reject, see (LEWONTIN and KRAKAUER 1973) for an early example. One approach for protein coding DNA sequences is to consider the relative rate of amino acid changing (non-synonymous) and amino acid preserving (synonymous) mutations, see for example (LI *et al.* 1985; MIYATA and YASUNAGA 1980). If the changes occur at an equal rate, a change that alters an amino acid is as likely as a change that does not. Since the amino acids build proteins and it is the action of proteins that is exposed to natural selection, this is regarded as evidence for the neutral hypothesis. If the rate of non-synonymous change is lower than the rate of synonymous change, mutations that change the amino acid are being fixed in a population at a lower rate than mutations that do not. It is argued that this is a sign of stabilising, or negative selection. Mutations that change the amino acid incur a survival cost and hence are eradicated from the population by selection. If the rate of non-synonymous change is higher than the rate of synonymous change, mutations that change the amino acid are being fixed in a population at a higher rate than mutations that do not. It is argued that this is a sign of adaptive evolution, also known as positive selection, because mutations that change the amino acid confer some sort of survival benefit and hence are become prevalent. Methods for detecting the rate of non-synonymous and synonymous changes will be discussed in a later section and the different performance of different methods form a second strand running through the results section.

CHAPTER 2. METHODS AND THEIR APPLICATION

For the purposes of this thesis, we divide the approaches used to address the problems that were described in the last chapter into two categories. The first approach, described here as the cladistic approach, reasons about evolution within a cladistic framework. The second approach, termed here statistical phylogenetics, regards evolution as a statistical process and approaches the problem using the standard techniques of statistical inference. In this section we describe the principles behind each approach and the way they are applied to the biological problems described, before going on to describe previous comparisons between them.

2.1 The cladistic approach

In early studies, see (ANDREWS 1904) for an example, evolutionary relationships were constructed in an ad hoc manner, relationships being based on small groups of shared characters. One systematic method, attributed to (HENNIG 1966), is the cladistic method.

The cladistic method rests on an assumption that the relationship between organisms can be represented by a tree that describes the interrelationships of clades. A clade is a subset of organisms that all share a common ancestor that is not shared by any organsim outside the subset. Thus humans form a clade within primates which form a clade within the mammals which in turn form a clade within the vertebrates. One of the purposes of the cladistic method is to find the topology of the tree. To do this one proceeds by identifying the taxa that are of interest and determining which characters of

those taxa are to be included in the analysis. For example when constructing the evolutionary relationships of the jawless vertebrates, the characters of interest will include the presence or absence of seven or more paired gill pouches, a light-sensitive pineal eye and the position of the branchial arches (JANVIER 1996). The taxa being studied are then scored for the presence or absence of the characters. The cladistic analysis proceeds by reconstructing the ancestral nodes on the tree, or cladogram, by minimising the number of derived changes or apomorphies. A cladogram with fewer derived changes is preferred to one with more derived changes. Derived changes shared by members of a clade are termed synapomorphies, derived changes that have occurred in different clades are termed homoplasies. For example the 2 holed skulls of diapsids (BENTON 1985) are a synapomorphy but the wing of bats, birds and pterosaurs are a homoplasy. The fundamental principle behind cladistics is that it is similarities due to derived changes that give information about evolutionary relationships. In contrast similarities due to a character inherited from a basal ancestor, or plesiomorphies, do not give information about evolutionary relationships. One possible justification given (KLUGE 2001) is that in minimising the occurrence of derived changes, one obeys Occam's Razor by not unnecessarily multiplying entities where there is no need.

2.1.1 Application to phylogeny reconstruction

Phylogeny reconstruction in the cladistic framework involves constructing a tree that minimises the number of observed derived changes. Early work, for example (KLUGE and FARRIS 1969), was done on morphological traits, in an effort to devise a method such that different investigators, given the same initial data, would reach the same conclusions. In an early morphological simulation study (CAMIN and SOKAL 1965), it was argued that

the best means of reconstructing hypothetical organisms was to choose the tree that minimised the number of derived changes that were necessary to explain the morphological differences observed at the tips. Various schemes that make different assumptions about the course of evolution have been suggested. The Camin-Sokal scheme does not allow reversals, once a character has been derived it does not revert to its ancestral condition. The Dollo scheme (LEQUESNE 1974) does not allow independent gains of a properly specified derived condition. The Wagner parsimony scheme (FARRIS 1970) defines intermediate states for a character and when counting derived changes, infers passage through the intermediate states. However the simplest scheme is known as Fitch parsimony (FITCH 1971) though originally suggested in (KLUGE and FARRIS 1969) in which all changes are possible and count equally. We describe the Fitch algorithm in Box 1, following (DURBIN 1998).

Initialisation

Set $C = 0$; k = index of the root node.

Recursion

If k is a leaf node

Set $R_k = x_k$

Else if k is not a leaf node

Compute R_i, R_j for the daughter nodes i, j

If there are states present in both R_i and R_j, $R_k = R_i \cap R_j$

Else if no states are present in both R_i and R_j, $R_k = R_i \cup R_j$; increment C.

Termination

The least number of changes that can account for the states observed at the tips is given by C.

Box 1. The Fitch parsimony algorithm. The possible states at node i are denoted R_i and the total number of changes across the tree is denoted C. The data observed at tip j is denoted x_j. The algorithm begins at the root and recursively passes up the tree, terminating at the tips.

Ideally a minimal change score would be calculated for each tree and the tree that explained the data with the least number of changes would be chosen as the estimate of the true tree. In practice the number of possible trees grows very large as taxa are added (KLUGE and FARRIS 1969), hence heuristic methods are used to search through the tree space.

The application of Fitch parsimony to molecular sequences is straightforward in that the possible state space is either given by the possible bases present along the DNA molecule i.e. one of {T,C, A, G}, or by possible amino acids present along a peptide chain.

2.1.2 Application to modelling adaptive evolution

We discuss in this section two methods that rely on an inferred count of changes to detect adaptive evolution. One detects adaptive evolution between two sequences (NEI and GOJOBORI 1986) and the other at a single site over a phylogeny (SUZUKI and GOJOBORI 1999). They are included in this section because both methods rely on a count of differences between codons that is inferred under a minimum change criterion.

The first method was developed as a means of comparing the number of synonymous changes and non-synonymous changes that have occurred between two sequences. The approach taken is to infer from the sequences a count of the number of synonymous and non-synonymous changes that have occurred. This is then compared to the number of changes that could have been made given the state of the sequences. This latter number is referred to as the number of *available* changes. If inferred synonymous

and non-synonymous changes occur about as often as they would be expected to from the count of available changes then the hypothesis of neutral evolution cannot be rejected.

When performing the count of inferred changes, if a codon on one sequence differs by only one nucleotide from the codon on the other, the count of synonymous changes is incremented by one if the change is amino acid preserving and the count of non-synonymous changes is incremented by one if the change was amino acid altering. If there are two changes then there are two possible minimal pathways that account for the difference. If, for example, a codon TTT on one sequence has evolved into a codon GTA on the other these pathways are TTT (Phe) T GTT (Val) T GTA (Val) and TTT (Phe) T TTA (Leu) T GTA (Val). The first pathway involves one synonymous change and one nonsynonymous change whereas the second pathway involves two nonsynonymous changes. Under assumptions similar to those of Fitch parsimony, the changes along the minimal pathways are regarded as being equally likely. Hence the probability of the codon evolving along the first pathway is regarded as being equal to the probability of the codon evolving along the second. The count of synonymous substitutions is given by the sum over all possible pathways of the probability of a particular pathway being taken multiplied by the number of synonymous changes made along that pathway. In the case described this equals (0.5 x 1) + (0.5 x 0) = 0.5. Equivalently, the count of non-synonymous changes is given by the sum over the possible pathways of the probability of a particular pathway being taken multiplied by the number of non-synonymous changes made along that pathway, which in the case described equals (0.5 x 1) + (0.5 x 2) = 1.5. When there are three codon differences there are six possible pathways and the count proceeds in a similar way. Again each pathway is regarded as equally likely.

The counts of *inferred observations* of synonymous (s) and non-synonymous (n) changes are then compared to the number of *available* synonymous (S) and non-synonymous (N) changes given the state of the codon. The count of available synonymous substitutions is arrived at by taking each sequence and calculating the number of possible nucleotide changes that will not alter a codon. Once this is done for each sequence, an average is taken to get the mean number of available synonymous changes. Equivalently, the count of available non-synonymous substitutions is arrived at by taking each sequence and calculating the number of available changes that will alter the amino acid. The mean number of available non-synonymous changes is then calculated. There are two methods for comparing the expected number of changes and the inferred number of changes.

The first entails converting the number of inferred observations of synonymous and non synonymous changes (s and n) into a proportion of synonymous and non-synonymous differences, given by s/S and n/N respectively. This proportion is then converted to an estimate of the average number of synonymous (ds) and non-synonymous changes (dn) per site per unit time. This conversion is performed using a Jukes-Cantor correction, the details of which are described in the next section. The estimates so obtained are assumed to be normally distributed and a Z-test is performed to see whether dn is significantly different from ds. This method does take into account multiple changes. It could be argued therefore that perhaps it does not sit entirely within the cladistic framework. The method is described in this section because the count of inferred observations is made under the assumption of a minimal change pathway, with alternative pathways being weighted equally. This seems to be similar to the minimum-change assumptions made under Fitch parsimony. The Nei-Gojobori method might be more properly regarded as a mixed

cladistic/statistical one. However it would seem that if the cladistic-style assumptions underlying the original count are incorrect, then the statistical correction will perform poorly. Indeed it has been shown (YANG and NIELSEN 2000) that this is the case. When codon frequencies are biased alternative pathways are not equally likely and the method performs poorly.

The second method for making the comparison (ZHANG *et al.* 1998) takes the count of inferred observations and compares them directly with the count of available synonymous and non-synonymous changes using a Fisher exact test. This seems to be a cladistic approach, and the method is used in conjunction with parsimony reconstruction when analysing adaptive evolution on ancestral branches (ZHANG *et al.* 1998).

The other method we deal with in the thesis involves detecting adaptive evolution at a particular codon in a sequence (SUZUKI and GOJOBORI 1999). The method requires many sequences, with known evolutionary relationships. Initially, ancestral sequences are reconstructed using parsimony or statistical methods (YANG *et al.* 1995b). Then the number of synonymous and non-synonymous changes that have occurred over the phylogeny at a particular codon site is counted using the ancestral sequences as real data points. The count is then made in the same way as in the previous method, weighting possible alternative pathways equally. The number of available changes is calculated by calculating the number of changes possible at each ancestral state and taking an average weighted by branch length. Where more than one change separates two codons at different ends of a branch, the available changes at the intermediate states are taken into account. The proportion of inferred synonymous changes is compared to the proportion of available synonymous changes using a binomial distribution.

It can be seen that this method sits squarely within the cladistic framework. The minimum number of changes that can explain the data is regarded as a true count of changes. Alternative pathways are equally weighted under the same assumptions as those of the Fitch parsimony scheme. It would seem that, as with the previous method, if the minimal count of changes is incorrect then the method will perform poorly. The question of inaccurate counts will be addressed in chapter 5 of the thesis.

2.2 Statistical phylogenetics

When taking a statistical approach to phylogenetics a researcher regards phylogenetics as any other statistical problem. A probabilistic model of sequence evolution is defined and its parameters are estimated using standard statistical techniques. Probabilistic models of sequence evolution are defined as time-homogenous Markov processes (PAPOULIS 1984) with a finite number of states in continuous time. Some of the parameters of the probabilistic model define an instantaneous rate matrix, denoted **Q**, that describes the rate of change of the probability that a site is in a given state. Once **Q** has been defined, the Kolmogorov equations can be used to calculate the probability of observing the data. In this section we describe the models of sequence evolution that we use in this thesis and how they are applied when there are only two sequences. We then proceed to explain how they are applied to the many sequence case. We then discuss the statistical methods of parameter estimation and hypothesis testing that are used and how they have been applied to the problems of evolutionary biology that have already been described.

2.2.1 The two sequence case

The first application of the Markov process framework to molecular evolution was given by Jukes and Cantor (JUKES 1969). In the Jukes-Cantor model, possible states in a DNA sequence are given by the possible bases at a site {T, C, A, G}. We write a site in state i as x_i. The Jukes-Cantor model provides a way to calculate $P(x_j \mid x_i ; b)$, where b is the distance measured in average number of changes per site. For a pair of sequences we

are interested in calculating $P(x_i, x_j; b) = P(x_i \mid x_j; b)P(x_j)$. To do this, the instantaneous instantaneous rate matrix for the Jukes-Cantor model is defined as:

$$\frac{\partial P(b)}{\partial b} = P(b)\mathbf{Q} = P(b)\begin{bmatrix} -3 & 1 & 1 & 1 \\ 1 & -3 & 1 & 1 \\ 1 & 1 & -3 & 1 \\ 1 & 1 & 1 & -3 \end{bmatrix} \quad \text{.....(2.1)}$$

The resulting Kolmogorov forward equation is then solved to give:

$$P(b) = P(0)e^{\mathbf{Q}b} \quad \text{.....(2.2)}$$

In the Jukes-Cantor model, the equilibrium probability of observing a DNA site in state x_i is 0.25, ie all states are equally likely. Hence the probability of observing the state x_i at one end of a branch and state x_j at the other is given by:

$$P(x_i \mid x_j; b)P(x_j) = \begin{cases} \frac{1}{16}(1 + 3e^{-4b}), \text{i} = \text{j} \\ \frac{1}{16}(1 - e^{-4b}), \text{i} \neq \text{j} \end{cases} \quad \text{.....(2.3)}$$

A sequence consists of many sites, so when considering two evolutionarily related sequences we observe many pairs $\{x_i, x_j\}$. In the Jukes Cantor model sites are considered to be evolving independently from each other. Hence writing the s^{th} such pair in a sequence of length n bases as $\mathbf{x}(s)$, we have:

$$P(\mathbf{x}) = \prod_{s} P(\mathbf{x}(s)) \quad (2.4)$$

Where P(**x**) is the probability of observing all the pairs $\{x_i, x_j\}$. In the standard statistical framework, each item of data is the pair $\{x_i, x_j\}$ and the probabilities $P(x_i, x_j)$ can be regarded as the probabilities defining a multinomial distribution (HUELSENBECK and CRANDALL 1997). Each pair $\{x_i, x_j\}$ can be considered the outcome of a trial in which a site on sequence j is picked at random from its equilibrium distribution and allowed to evolve over distance b. Its state is then observed on sequence i. In general the evolutionary distance, b, is not known and needs to be estimated.

Discussion of the method by which the parameter b is estimated is deferred to a later section. Additionally to address the problem of phylogeny reconstruction we have to extend our calculation of equation (2.4) from the two sequence to the many sequence case. However, before doing this we describe some extensions to the Jukes Cantor model.

The limitations of the Jukes-Cantor model can be seen immediately. All changes between states are regarded as equally likely. No account is made for transition-transversion bias or base usage frequencies that differ from equality. We have seen in chapter 1, that the process of DNA replication is relatively well understood and some mutations are inherently more likely than others. In this section we address methods that address these limitations.

A model that accounts for transition/transversion bias is due to Kimura (KIMURA 1980). The instantaneous rate matrix is equivalent to:

$$\frac{\partial P(b)}{\partial b} = P(b)\mathbf{Q} = P(b)\begin{bmatrix} -\kappa-2 & \kappa & 1 & 1 \\ \kappa & -\kappa-2 & 1 & 1 \\ 1 & 1 & -\kappa-2 & \kappa \\ 1 & 1 & \kappa & -\kappa-2 \end{bmatrix} \quad \text{.....(2.5)}$$

Calculation of the probability of the pair $\{x_i, x_j\}$ is performed in the same manner as before, by solving the Kolmogorov forward equation. There are two parameters to be estimated, b as before and the transition/transversion rate ratio, κ.

A model that accounts for base usage bias is due to Felsenstein (FELSENSTEIN 1981; FELSENSTEIN 1985). The instantaneous rate matrix is given by:

$$\frac{\partial P(b)}{\partial b} = P(b)\mathbf{Q} = P(b)\begin{bmatrix} -\pi_C-\pi_A-\pi_G & \pi_C & \pi_A & \pi_G \\ \pi_T & -\pi_T-\pi_A-\pi_G & \pi_A & \pi_G \\ \pi_T & \pi_C & -\pi_T-\pi_C-\pi_G & \pi_G \\ \pi_T & \pi_C & \pi_A & -\pi_T-\pi_C-\pi_A \end{bmatrix}$$
$$\text{.....(2.6)}$$

Where π_i is the equilbrium probability of base i. Calculation of the probability of observing the pair $\{x_i, x_j\}$ proceeds as before. This time it can be seen the there are 4 parameters to estimate: b and three of the four base equilibrium parameters.

These two models are synthesised into a model that accounts for both transition transversion bias and base usage patterns, the HKY85 model described by the instantaneous rate matrix (HASEGAWA *et al.* 1985):

$$\frac{\partial P(b)}{\partial b} = P(b)\mathbf{Q} = P(b)\begin{bmatrix} -\kappa\pi_C - \pi_A - \pi_G & \kappa\pi_C & \pi_A & \pi_G \\ \kappa\pi_T & -\kappa\pi_T - \pi_A - \pi_G & \pi_A & \pi_G \\ \pi_T & \pi_C & -\pi_T - \pi_C - \kappa\pi_G & \kappa\pi_G \\ \pi_T & \pi_C & \kappa\pi_A & -\pi_T - \pi_C - \kappa\pi_A \end{bmatrix}$$

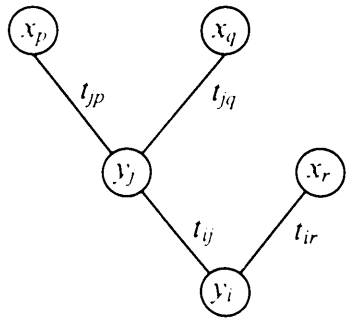
.....(2.7)

In this case there are 5 parameters to estimate.

It can be seen that Kimura's model is a special case of HKY85 as is Felsenstein's. Jukes-Cantor is a special case of Kimura's and Felsenstein's. There exist further generalisations of HKY85. One adds an extra parameter so that the transition rate between purines is different from the transition rate between pyrimidines (TAMURA and NEI 1993). A further generalisation is the General Time Reversible model which is the most general possible symmetrical instantaneous rate matrix. The logical conclusion of the generalisation process is the general irreversible model which has a different parameter for every change x_i T x_j.(YANG 1994a) In this thesis the most general matrix used is the HKY85. It has been argued from simulation studies (YANG 1994a) that this provides a useful compromise between the number of parameters and biological realism.

2.2.2 The many sequence case

So far we have discussed the two sequence case. To consider how we calculate the probability in the many sequence case we consider the following diagram.



This diagram represents a single site in the sequence. The state of the site on the p^{th} sequence is given by x_p, the state of the site on the q^{th} sequence is given by x_q and the state of the site on the r^{th} sequence is given by x_r. The sequences x are joined by a phylogeny. The state of the site at the j^{th} node is described by y_j. y_j is unknown. Equivalently the state of the site on the i^{th} root sequence is described by y_i. The branch lengths, or distances between nodes, are marked on the diagram and referred to collectively as **t**. It can be seen that if y_i and y_j were known, the probability of observing the data would be the probability of a site changing from state y_i T y_j in time t_{ij}, changing from state y_i to x_r in time t_{ir},changing from state y_j to x_p in time t_{jp} and finally changing from state y_j to state x_q in time t_{jq}. Since these changes occur independently we would write the probability of the site starting in state y_i and evolving down the tree to states x_p, x_q and x_r as:

$$P(x_p, x_q, x_r, y_i, y_j; \mathbf{t}, \theta) = P(x_p | y_i; t_{ip}, \theta).P(x_q | y_j; t_{jq}, \theta).P(y_j | y_i; t_{ij}, \theta).P(x_r | y_i; t_{ir}, \theta)$$

.....(2.8)

where θ represents the parameters of the instantaneous rate matrix.

Unfortunately we do not know the sequence at the internal nodes. To get around this problem we sum over all possible unknown states in the following manner:

$$P(x_p, x_q, x_r; \underline{t}, \theta) = \sum_{y_j} P(x_p | y_j; t_{jp}, \theta).P(x_q | y_j; t_{jq}, \theta).\sum_{y_i} P(y_j | y_i; t_{ij}, \theta).P(x_r | y_i; t_{ir}, \theta)$$

.....(2.9)

Thus we can calculate the probability of observing $P(x_p, x_q, x_r; \mathbf{t}, \theta)$ at a single site. Equation (2.9) forms the basis of Felsenstein's tree pruning algorithm (FELSENSTEIN 1981) that steps back from the tips summing over unknown states at internal nodes. This algorithm in turn forms the basis of the algorithm described in the nth chapter.

2.2.3 Adding biological realism

In the models so far described, each site evolves at a constant rate. Thus the probability of two sites which are in the same state at the start of a branch changing over that branch is the same wherever the sites are in a biological sequence. Biologically this is unrealistic because some sites will code for amino-acids that form a part of a protein essential to its catalytic function. Alternatively some sites will code for an amino-acid that is purely structural, the changing of which will provide little change in the proteins

function. It can be seen that the first kind of change will be in general selected against and hence be observed to happen at a slow rate but that the second will happen more frequently.

It has been shown (CHANG 1996) that likelihood methods for phylogeny reconstruction can be inconsistent if there is a large proportion of invariant sites that cannot change state and if the model assumes one rate for all sites. These sites shorten the branch length artificially, hence the true branch length along which the variable sites are evolving can never be reconstructed and this makes the method inconsistent.

A model that accounts for this variation in rate was given by Yang (YANG 1993; YANG 1994b). The branch lengths calculated in equation (2.9) are scaled by a factor L. The value of L is described by a gamma distribution with parameters a, $1/a$ such that its mean is 1.0. To give an example the probability of observing the data x_p, x_q, x_r in figure 2.1 at a site is given by:

$$P(x_p, x_q, x_r; \mathbf{t}, \theta, a) =$$
$$\sum_{L}\sum_{y_i} P(x_p \mid y_j; Lt_{ip}, \theta).P(x_q \mid y_j; Lt_{jq}, \theta)\sum_{y_i} P(y_j \mid y_i; Lt_{ij}, \theta).P(x_r \mid y_i; Lt_{ir}, \theta)P(L; a)$$

.....(2.10)

Here $P(L; a)$ is the probability of drawing the value L from a gamma distribution with parameters a, $1/a$. . This model of between-site variation will be contrasted with the model of heterotachous evolution to be described in chapter 3.

Statistical methods are also used to detect adaptive evolution in a similar way. In these methods the unit of evolution is not the DNA base but the codon triplet. The instantaneous rate matrix for a single site is given by:

$$q_{ij} = \begin{cases} 0 & \text{If codon } i \text{ and } j \text{ differ by more than 1 nucleotide} \\ \pi_j & \text{If codon } i \text{ and } j \text{ are synonymous and differ by 1 transversion} \\ \kappa\pi_j & \text{If codon } i \text{ and } j \text{ are synonymous and differ by 1 transition} \\ \omega\pi_j & \text{If codon } i \text{ and } j \text{ are nonsynonymous and differ by 1 transversion} \\ \omega\kappa\pi_j & \text{If codon } i \text{ and } j \text{ are nonsynonymous and differ by 1 transition} \end{cases}$$

.....(2.11)

Here ω models the ratio of the rate of amino-acid changing mutations to non amino-acid changing mutations. Under Kimura's hypothesis of neutral evolution the ratio of these rates should be 1, amino-acid changing mutations should happen as often as non-amino acid changing mutations. In the simple form of this model each site has the same value of ω. The neutral hypothesis has the value of ω fixed at 1, under the alternative hypothesis ω is allowed to vary. Thus the null is nested within the alternative hypothesis. In chapter 3 it is this simple form of the model that is studied. More complex models exist (NIELSEN and YANG 1998) that all have a distribution describing values of $\omega <= 1$ nested within a distribution that has an extra class describing values of $\omega > 1$. More complex models still, allow the ω value at a site to change on a branch (ZHANG *et al.* 2005). However, because of their high dimensionality, these more complex models are not investigated in this thesis.

2.2.4 Application of statistical inference to evolutionary biology

So far we have described how the probability of observing the data is calculated given that we know the parameters and the evolutionary relationships between the species. In practice we do not know either. The parameters and relationships have to be estimated from the data. In this section we describe the relevant statistical methods for doing this.

The first method we describe is that of maximum likelihood. Writing the probability of observing the data, **x**, given the parameter θ, as $P(\mathbf{x} ; \theta)$, we choose as our estimate of the parameter, $\hat{\theta}$, the value that maximises:

$$\ell(\hat{\theta}) = \arg\max_{\theta} \log[P(\mathbf{x};\theta)] \quad (2.12)$$

In situations where models are nested, such as the DNA sequence evolution models and codon models already described, we expect some increase in probability when we calculate the probability of the data under the more general model even when the nested model is true. The distribution of the increase in the log of the probability is given by (BARNDORFF-NIELSEN 1994) :

$$2[\ell(\hat{\theta}) - \ell(\hat{\theta}_0)] \sim \chi^2_k \quad (2.13)$$

Where $\hat{\theta}_0$ is the maximum likelihood estimate under the true nested model, $\hat{\theta}$ is the maximum likelihood estimate under the more general model and k is the number of extra

parameters in the more general model. If the difference in probability is greater than that expected from the χ^2 distribution, the nested model can be rejected.

When reconstructing a phylogeny, a tree topology (T) is chosen and the method of maximum likelihood is used to estimate the unknown parameters, in this case the evolutionary distances between nodes (the branch lengths, **t**) and the parameters of the rate matrix, θ. The most general application of this approach would be to have a different instantaneous rate matrix for each branch (BARRY 1987). In practice the rate matrix is kept constant across the tree in most cases, though this may cause problems in very deep phylogenies where different animal phyla have different base usage patterns (RUIZ-TRILLO *et al.* 2002).

Once the parameters, **t** and θ, have been chosen to maximise the probability of observing the data we have, for a given tree topology, the maximum likelihood of observing the data. We are then in a position to compare that likelihood with the likelihood of observing the data under the hypothesis of a different tree topology. It is defensible to choose the phylogeny with the highest likelihood of observing the data as the best supported hypothesis (FELSENSTEIN 1981). In practice we again run into the problem of a high number of possible trees and heuristic methods are used to search through the tree space (LEMMON and MILINKOVITCH 2002); (GUINDON and GASCUEL 2003). Different evolutionary trees represent different non-nested statistical models that each define a multinomial distribution. Direct comparison between the likelihoods can be performed, even though the models are non-nested, because all trees define models with the same number of outcomes. Thus the multinomial constant is the same in each case (YANG *et al.* 1995a).

The method of maximum likelihood is contrasted with Bayesian approaches. These calculate the explicit probability of observing a parameter value θ using the formula:

$$P(\theta \mid \mathbf{x}) = \frac{P(\mathbf{x} \mid \theta)P(\theta)}{\sum_{\theta} P(\mathbf{x} \mid \theta)P(\theta)} \quad \text{.....(2.14)}$$

It can be seen that this method relies on the existence of assigning to each particular value of θ a prior probability, $P(\theta)$. This can either be based on a belief that exists before the data is collected or the prior probabilites of all possible values of θ can be said to be equal. It is argued (JEFFREYS 1961) that this latter approach is a precise way of saying that there is no ground for choosing between the alternatives. However it has also been argued that assigning an explicit probability to a value of θ in the absence of any information is incorrect (EDWARDS 1992).

The Gamma model of rates is an example of an empirical Bayesian model. The gamma distribution is an expression of prior belief in the distribution of the parameter of interest, in the notation of equation (2.10), L. The model is empirical, in that the form of the prior is derived from the data (OWEN 2001) by means of the method of maximum likelihood. The model is Bayesian in that the choice of a gamma distribution to describe the distribution of rates is based on a prior belief that because of its flexibility of shape it will be a useful model whatever the truth, rather than being derived from an underlying biological process.

2.3 Comparisons of cladistic and statistical methods

In this section we describe previous work comparing statistical and parsimony approaches to both tree reconstruction and the study of adaptive evolution.

2.3.1 Tree reconstruction

The comparisons of parsimony and likelihood methods are almost as old as the methods themselves. Cavalli-Sforza and Edwards in an early paper (CAVALLI-SFORZA 1967) suggest three methods for tree estimation. One of these is a minimum evolution method that uses the idea that a plausible estimate of the true tree is given by the tree that invokes the minimum total amount of evolution. This idea is similar to that of the cladistic methods. Additionally they propose a likelihood model that explicitly models evolution probabilistically. Their likelihood model is different from more modern methods in that they use continuous allele frequencies as their data and attempt to simultaneously estimate the allele frequencies at ancestral nodes. They found that both methods performed tolerably, but they argued that the success of their minimum evolution tree was because of its closeness to the maximum likelihood tree. Arguably this is still a defensible view.

An attempt to derive a probabilistic interpretation of parsimony was made by (FARRIS 1973). In his model he defined an evolutionary hypothesis as including both the phylogenetic tree itself and the state of the system at a given point on the tree. Under fairly general conditions he showed that replacing any two changes by one change must increase the probability of the evolutionary hypothesis when it was so defined. Thus to maximise the probability of observing the data one chose the evolutionary hypothesis with the

smallest number of changes. In other words one chose the tree that would be inferred under the assumptions of parsimony.

The problems with both this method and the likelihood method of Cavalli-Sforza and Edwards were addressed by (THOMPSON 1975). She addressed the issue of consistency. In both models, and hence by extension parsimony methods, an attempt is made to estimate the state at ancestral nodes. These states are treated as a parameter to be estimated. Hence by adding data, one adds an extra set of parameters to be estimated at the nodes. Maximum likelihood is not necessarily consistent if there are more parameters than there are data. In order to make the method consistent, it is necessary to sum over the unknown states at the nodes.

The issue of consistency was addressed by (FELSENSTEIN 1978) who showed that under certain circumstances parsimony was indeed inconsistent. In a four taxon tree, parsimony became inconsistent when many changes had occurred on two branches in different clades but not on the rest of the tree. This became labelled the "long-branch attraction" problem and the region of parameter space in which it occurred became labelled the "Felsenstein zone". This was addressed by parsimony proponents who claimed a "Farris zone" (SIDDALL 1998) in which parsimony had a higher probability of reconstructing the true tree than did likelihood. The existence of the Farris zone is beyond doubt, as we will see in a later chapter, see also (YANG 1996). However likelihood is still consistent within this region.

(FELSENSTEIN 1983) proposed that parsimony was equivalent to likelihood when evolutionary change is rare. For small numbers of taxa, the probability of a site evolving data that was parsimony informative could be calculated. Thus formulas for the probability

of reconstructing the correct tree using parsimony methods could be devised (TAKEZAKI and NEI 1994; ZHARKIKH and LI 1993). These provided more evidence that in some circumstances parsimony was inconsistent and in some circumstances outperformed likelihood.

(GOLDMAN 1990) provided a description of parsimony similar to (FARRIS 1973), but limited to the two state case. He was able to show that a maximum likelihood method in which both the tree shape and ancestral states at the nodes were chosen to maximise the probability of reconstructing the data was equivalent to parsimony, provided the probability of change on a branch was not too great. Thus it was argued that parsimony analyses had at root a maximum likelihood justification, as argued by (CAVALLI-SFORZA 1967). However the particular form of the likelihood analysis, maximising a larger number of parameters than there is data, means that the likelihood method becomes inconsistent. It was also argued by Goldman that parsimony methods will coincide with likelihood methods when the rate of evolution at a site is small relative to the time that a sequence has been evolving.

This analysis seems to be contradicted by (HOLMES 2003) who argues that rather than a likelihood method with an super abundance of parameters, parsimony is best viewed as a non-parametric method. However rather than implying that parsimony has no parameters, she argues that this means parsimony is based on optimising potentially infinite-dimensional criteria. Whilst this is consistent with the previous work, it is unclear that a potentially infinite dimensional probability model as described by Farris and Goldman is necessarily equivalent to a nonparametric method. For example consistent likelihood methods exist for the estimation of parameters of interest defined by explicit

probability models even though infinitely many nuisance parameters are present, see for example (AMARI 1987).

We have seen that the definition of parsimony as "likelihood with an infinite number of parameters" has a long history. However the formulations so far presented regard the infinite number of parameters as arising from a simultaneous estimation of topology and state at the ancestral node. In an alternative formulation (STEEL and PENNY 2000; TUFFLEY and STEEL 1997), each site is associated with its own set of branch lengths. This is again an infinite parameter model. When the maximum likelihood method is used to reconstruct both the branch lengths and the topology, the resulting best tree is equivalent to the tree reconstructed by parsimony. This description has an intuitive appeal since the processes driving the selection of specific mutations are regarded as extremely complex. It is argued that complex processes must necessarily be described with many parameters. Therefore, the argument runs, since natural selection is such a complex process, an infinite parameter model must be better than a model with a low number of parameters.

The essential idea behind the method is that when there is one possible set of branch lengths per site, the probability of observing the data is maximised by: (1) collapsing some branch lengths to zero making them "no change" branches, and then (2) choosing a branch length that maximises the probability of observing one change for the other branches. This makes them "change branches". The likelihood will then be maximised by collapsing the tree in such a way as to maximise the number of "no change branches", which do not reduce the likelihood, and the minimise number of "change branches", which do. This produces a result equivalent to parsimony. It is worth noting that the result only holds for the case where all changes between states are equally likely. Also if we consider the Fitch

algorithm described, it can be seen that under parsimony, every internal node can be connected to a tip by branches along which no change has occurred. Thus if all ancestral nodes have a "no change" pathway to a tip, under the Tuffley-Steel scheme this pathway will have length zero and hence the state at the ancestral node will be the same as that at the tip. Effectively the ancestral state will have been reconstructed in the likelihood maximisation. This indicates a link between the Tuffley-Steel scheme and previous work.

It seems the case that, when applied to tree reconstruction, parsimony has a developed interpretation as a maximum likelihood estimate of an evolutionary history that includes both the phylogeny and the state of the ancestral nodes. The problem with this interpretation from a statistical perspective is that simultaneous estimation of the tree and the state at the nodes leads to a problem of more parameters than data.

2.3.2 Comparisons of methods for detecting adaptive evolution

In contrast to the problem of reconstructing evolutionary histories, cladistic approaches to detecting adaptive evolution have not been interpreted in a probabilisitic way. Comparisons have been made using simulation studies, for example (WONG *et al.* 2004), which seem to indicate that empirical Bayesian methods have more power than the method of (SUZUKI and GOJOBORI 1999). Additionally it has been shown that a minimum change method found a non-synonymous/synonymous rate ratio greater than one across a gene on a single branch in a phylogeny(MESSIER and STEWART 1997), though likelihood methods suggested that the null model could not be rejected (YANG 1998a). However a case has been presented that shows that in certain circumstances empirical Bayesian methods will assign a high probability of adaptive evolution to sites that are invariable

(ZHANG 2004). This problem has been addressed recently by making the empirical Bayesian method less empirical by introducing a hierarchy of priors to be integrated over (YANG *et al.* 2005).

To start to address the absence of theory, we develop methods in Chapter 5 for calculating the probability of a site making a given number of changes of interest over a phylogeny. Since cladistic methods often use a parsimony derived estimate of the number of changes, it is likely they will perform poorly if the estimate is wrong in a systematic way.

2.4 Notes on Implementation

In this section some implementation details are described that do not belong in other sections. Included here are some notes on optimisation and some notes on decisions made in the software design process. All software developed in the course of this project is available for download from http://www.ucl.ac.uk/~ucbpdcd/thesis/

2.4.1 Numerical optimisation methods

We have seen that the statistical methods described rely heavily on choosing parameters that maximise the likelihood of observing the data. In practice the equations describing the probability of observing the data can almost never be solved analytically, but see (YANG 2000a) for a case for which an analytical solution exists. Instead numerical methods are used to find the combination of parameter values that maximise the probability of observing the data.

To do this a numerical optimisation library was written that implements Newton's minimisation algorithm. This library takes a function and returns a vector of parameters for which the function is a minimum. Our maximisation problem is solved by writing the function in such a way that it returns -1 times the log likelihood, turning the maximisation problem into a minimisation. In practice it is not possible to calculate the gradient of the likelihood function with respect to all the parameters, though it is possible to calculate the gradient with respect to the branch lengths, see (GOLDMAN 1998) and (YANG 2000b). For the work in this thesis these methods were not found to make a large difference and quasi-Newton optimisation routines were used.

Writing the current estimate of the minimum as x_k, the step from that point p, the function to be minimised, F, can be approximated by taking a Taylor series expansion about the current point:

$$F(x_k + p) \approx F(x_k) + g_k{}^T p + \frac{1}{2} p^T G_k p \quad \text{.....(2.15)}$$

Here g_k is the gradient vector of F at the point x_k and G is the Hessian. The minimum of the right hand side of (2.15) will be achieved if p is a minimum of the quadratic function:

$$\phi(p) = g_k{}^T p + \frac{1}{2} p^T G_k p \quad \text{.....(2.16)}$$

A stationary point of (2.16) will satisfy:

$$G_k p_k = -g_k \quad \text{.....(2.17)}$$

The solution of (2.17) gives the Newton search direction (GILL 1981). Once a search direction has been found the algorithm proceeds by stepping along the search direction until it finds the minimum in that direction. The algorithm used in the optimisation library is the safeguarded polynomial interpolation method. Once the algorithm has stepped to a new estimate of the minimum, (2.17) is solved at the new point and the algorithm proceeds until a local minimum is found. Unfortunately in this case we cannot calculate $g(x)$ or $G(x)$. $g(x)$ is approximated using the forward difference formula. $G(x)$ is initiated as the identity matrix, then updated using the Broyden-Fletcher-Goldfarb-Shanno (BFGS) update that makes a first order approximation to G based on the change in gradient along the search direction.

CHAPTER 3. THE EFFECT OF SUBSTIUTION SATURATION ON LIKELIHOOD AND PARSIMONY ANALYSES.

3.1 Introduction

Homologous biological sequences that remain in evolutionary isolation from each other independently accumulate mutations. The more mutations accumulated, the more divergent the sequences. Attempts to make inferences about the evolutionary processes that have driven those mutations are limited if the divergence between those sequences becomes too great. This general concept is known as saturation (SMITH and SMITH 1996).

Counting methods involve inferring the number of changes that have occurred across a phylogeny and then deducing the pattern of evolutionary change from the inferred counts. The number of inferred changes is generally calculated using a minimum-change criterion. Thus it is argued that when reconstructing the pattern of past events, the best supported hypothesis of the evolutionary history is that with the minimal number of inferred changes. It can be seen that this minimum-change criterion no longer holds where there is high sequence divergence. Multiple changes can accumulate at a single site, leading to a low correspondence between the true number of changes and the number inferred under a minimum change criterion (PHILIPPE and LAURENT 1998). Thus inferences based on this criterion can be misleading. It has been argued (BROCCHIERI 2001) that even when high levels of sequence identity (40%-50%) are observed, saturation cannot be ruled out.

On the face of it, likelihood-based methods should not suffer from this problem. They are defined in terms of an instantaneous rate of change and since any evolutionary distance is, in one sense, made up of instants, the occurrence of multiple changes should be

accounted for in the likelihood calculation. Whilst this is true, large sequence divergence causes a different problem for the likelihood-based methods. As mutations accumulate at each site in the sequence, the probability that a site is in a given state approaches an equilibrium value that is independent of the state in which the site started (WHELAN *et al.* 2001). Thus if two sequences are highly divergent, the probability that a site is observed in a given state in one sequence state will approach an equilibrium value that is independent of its state in the other. This equilibrium probability will tell us little about the evolutionary process of interest.

High sequence divergence causes problems for both minimum-change analyses and likelihood analyses, but by a different mechanism in each case. Therefore we split the two mechanisms of saturation into saturation by criterion violation (S.C.V.) for minimum-change analyses and saturation by approach to an equilibrium (S.A.E.) for likelihood analyses. With this in mind we shall investigate how S.A.E. and S.C.V. affect inferences about the strength of selection in codon models and transition/transversion rate ratios in nucleotide models as well as extending previous work (YANG 1998b) to find how they affect phylogeny estimation.

The codon model of (GOLDMAN and YANG 1994; MUSE and GAUT 1994) has a single parameter (ω) that models the relative rate of non-synonymous (amino acid changing) and synonymous (amino acid preserving) nucleotide changes. In addition there is a transition/transversion rate ratio parameter (κ) that models the relative rate of transitions to transversions and a set of codon frequency parameters (Π) that define the equilibrium distribution. In the model we shall examine here, ω is held constant for each site in the sequence. This model has been used in the identification of pseudo genes (ZHENG *et al.*

2005), studying paralogous gene pairs (MAERE *et al.* 2005), olfactory receptors (GILAD *et al.* 2005) and the evolution of orphan genes, i.e. genes with no homologue in distantly related species (DOMAZET-LOSO and TAUTZ 2003). It has also been used in the comparative genomic analysis of the turkey and chicken (AXELSSON *et al.* 2005) as well as in the comparative analysis of prokaryotic genomes (CANBACK *et al.* 2004; FRIEDMAN *et al.* 2004). When considering pairs of sequences at high divergences the probability of observing a state in one sequence will tend to an equilibrium that is both independent of the state in the other sequences and independent of the value of the parameter of interest, ω. Thus the model has the conditions necessary for S.A.E.. If, as argued (BOFFELLI *et al.* 2004) comparative genomics moves to the extremes and the evolution of divergent sequences is studied more frequently, it is likely that analyses will be limited by S.A.E..

The parametric model of codon evolution is usually held in contradistinction to the method of Nei and Gojobori (NEI and GOJOBORI 1986), modified by Zhang et al (ZHANG *et al.* 1998) to take some account of the transition/transversion rate ratio. In this method, one counts the minimum number of synonymous and non-synonymous nucleotide changes that can account for the codon differences between sequences. In cases where two or more changes are needed to account for the observed codon difference, different minimal pathways are weighted equally. In the modified method (ZHANG *et al.* 1998) these inferred counts are treated as actual counts and compared directly to the possible numbers of non-synonymous or synonymous mutations. The theory behind this being that under neutral evolution the ratio of actual non-synonymous to actual synonymous changes should be equal to the ratio of possible non-synonymous to possible synonymous changes. The

numbers of possible changes are calculated taking into account the transition/transversion rate ratio of Kimura's model of nucleotide substitution (KIMURA 1980).

It can be seen that the model has the conditions necessary for S.C.V. since divergent sequences can build up many changes at a single site making the counts inaccurate. In the original method (NEI and GOJOBORI 1986) the inferred counts are changed to a rate and this conversion is supposed to account for many changes at a single site. However since the inferred number of differences is calculated using equally weighted minimal pathways it is possible that this method is also vulnerable to S.C.V.. These methods have been used widely, for example to investigate the evolution of prolactin in primates (WALLIS *et al.* 2005), in examining selection across the fungal genome (HUGHES and FRIEDMAN 2005), in the comparison of avian myostatin genes (GU *et al.* 2004) and in the evolution of hominoid-specific forms of neural genes (LI *et al.* 2004).

Saturation is also a confounding problem when the transition/transversion rate ratio is estimated. Transition mutations are changes at a nucleotide site that replace a purine with a purine and a pyrimidine with a pyrimidine, as opposed to transversions that interchange them. The transition/transversion rate ratio has been studied in the chicken genome (HILLIER *et al.* 2004) and is of interest when investigating the evolution of isochores (ARNDT *et al.* 2003). The likelihood model we examine here is Kimura's 2 parameter model (KIMURA 1980). The nucleotide equilibrium frequencies are each 0.25 and do not depend on the parameter of interest, the transition/transversion rate; hence it is a candidate for S.A.E.. This likelihood method is contrasted with Ina's unbiased estimate of counts of transitions and transversions (INA 1998).

Finally it has long been recognised that saturation becomes a problem in the reconstruction of phylogenies. This is the long branch attraction problem (FELSENSTEIN 1978), in which trees with long branches are reconstructed with long branches erroneously grouped in the same clade. Optimal sequence divergences have already been investigated (YANG 1998b); here we extend that work to investigate whether large sequence divergence makes phylogeny reconstruction impossible.

3.2 Methods

Likelihood analyses have the advantage of coming with a well-developed theory, see for example (EDWARDS 1992). Here we take the methods outlined in (GOLDMAN 1998) and apply them to the models of interest. In general, we have data **x** made up of n observations of patterns at n sites. We also have a probabilistic model determined by the true parameter(s) θ^*. The likelihood of the data is the probability of observing the data given the parameter value θ, written $P(\mathbf{x}; \theta)$. The log-likelihood is written as $\log[P(\mathbf{x} \mid \theta)]$. When inferring the value of θ^* from the data, we choose the estimate $\hat{\theta}$ that maximises the log-likelihood.

Starting with the familiar support function:

$$S(\theta) = \ell(\theta \mid \mathbf{x}) = \log[P(\mathbf{x} \mid \theta)] \quad \ldots\ldots(3.1)$$

where $P(\mathbf{x} \mid \theta)$ is the probability of having observed the data set **x** given parameter value θ. We choose as our estimate the value of θ that maximises S for the data we have observed.

Taking the Taylor expansion of S around $\hat{\theta}$, we get:

$$S(\theta) \approx S(\hat{\theta}) + (\hat{\theta} - \theta)^T \left.\frac{\partial S}{\partial \theta}\right|_{\hat{\theta}} + (\hat{\theta} - \theta)^T \left.\frac{\partial^2 S}{\partial \theta^2}\right|_{\hat{\theta}} (\hat{\theta} - \theta) \quad \ldots\ldots(3.2)$$

Where the differentials are evaluated at $\theta = \hat{\theta}$. However since we have chosen the estimate $\hat{\theta}$ that maximises S, it can be seen that $\partial S / \partial \theta$ at that point is zero. Thus in some sense the

difference in support between $\hat{\theta}$ and a θ in its immediate neighbourhood is given by minus the second derivative of the support function. Thus it is intuitive to use the second derivative as a measure of the information present in the data pertaining to the true value of the parameter. The larger the second derivative, the larger the difference in information between θ and points in its immediate neighbourhood. This is the observed information.

However if we know our parameters θ^*, and wish to know how much information we can expect from an experiment, we can define an expected support function that is equivalent to the observed support function above. Writing <X> for $E_{\theta^*}(X)$:

$$\langle S(\theta) \rangle = \sum_{\mathbf{x}} P(\mathbf{x} \mid \theta^*) \log[P(\mathbf{x} \mid \theta)] \quad \text{.....(3.3)}$$

Here P($\mathbf{x} \mid \theta$) is the probability of one possible outcome of an experiment and the summation is over all possible outcomes.

We can take an expansion of the expected support function about the known parameters:

$$\langle S(\theta) \rangle \approx \langle S(\hat{\theta}) \rangle + (\hat{\theta} - \theta)^T \left\langle \frac{\partial S}{\partial \theta} \right\rangle + (\hat{\theta} - \theta)^T \left\langle \frac{\partial^2 S}{\partial \theta^2} \right\rangle (\hat{\theta} - \theta) \quad \text{.....(3.4)}$$

Noting that the expected differentials are evaluated at $\theta = \theta^*$, we get an expected, or Fisher information analogous to the observed information:

$$\langle I(\theta) \rangle = -\left\langle \frac{\partial^2 S}{\partial \theta^2} \right\rangle \quad \text{.....(3.5)}$$

This is the curvature of the expected support function. This time the expected information gives us in a sense the expected difference between the expected maximum of

the support function, ie the real value θ^*, and its immediate neighbourhood. As a tool for experimental design, it allows us to ask the question, for our real value θ^*, how much information can we expect from our experiment?

In the two-sequence problem each pair of sites, one at each end of the branch, defines a possible outcome to the evolutionary experiment and hence a data point. We have a multinomial distribution whose probabilities are functions of the parameters in the transition rate matrix. Since each site is modelled as an independent trial, the expected information is additive, i.e. n sites contain n times as much information as 1 site.

The expected support function gives us one tool to address the problem of how long is too long. For a given branch length b^* and $^{dN}/_{dS}$ ratio ω^* or transition/transversion rate ratio , κ^*, we can calculate the expected information present in n sites. By increasing b^* we can plot how the information falls for a given ω^* or κ^*. Eventually there will come a point when the branch length becomes too long and no information is present. We have restated the question about saturation as: how much information is too little information?
To answer this we use the asymptotic result (EDWARDS 1992):

$$\langle I(\theta)\rangle^{\frac{1}{2}}(\hat{\theta}-\theta^*) \sim N(0,\mathbf{i}) \qquad (3.6)$$

Here $\mathbf{i}$ is the identity matrix, and $N(0, \mathbf{i})$ is the spherical normal distribution with unit variance.

Hence, whether we are interested in the non-synonymous/synonymous rate ratio, or the transition/transversion rate ratio, for a known pair of parameters $\{\omega^*, b^*\}$ or $\{\kappa^*, b^*\}$, we can calculate the expected information matrix. The inverse of this will be the variance

of a normal distribution which will be centred at the actual parameter values and will define the distribution of estimates $\{\hat{\omega}, \hat{b}\}$ and $\{\hat{\kappa}, \hat{b}\}$. Whichever parameter we are interested in, either ω or κ, as the actual value of the branch length, b^*, gets very large the variance of the estimate $\hat{\omega}$ or $\hat{\kappa}$ becomes large. At some point we will decide that the probability of our estimate being within a reasonable distance of the actual value is so small that it is not worth proceeding with the analysis.

It is worth noting here that we are assuming that the samples are large enough for the asymptotic result to hold.

Additionally we have the standard result:

$$2\left[\log P(x \mid \hat{\theta}) - \log P(x \mid \theta_0)\right] \sim \chi_d^2 \quad \text{.....(3.7)}$$

θ_0 is the maximum likelihood estimate under a true model with m parameters, and $\hat{\theta}$ is the maximum likelihood estimate of a model with $m+d$ parameters that has the true model embedded in it. The m-parameter model is a special case of the $m+d$ parameter model. If twice the log-likelihood difference is larger than we would reasonably expect from the Chi-squared distribution of the statistic, we reject the hypothesis that the more constrained model adequately explains the data.

We applied these two standard results to our two likelihood models of interest. Firstly we studied what happens to the predicted variance of the estimate of the nonsynonymous/synonymous rate ratio as we increase the divergence between 2 sequences. For the simulation at various divergences, we used the simple single rate model for the non-

synonymous/synonymous rate ratio with one value of ω. The transition/transversion rate ratio was fixed at 5.0 and the codon frequencies fixed at equality. The value of ω and the divergence were then estimated from the simulated data, again with the transition/transversion rate ratio fixed at 5.0 and the codon frequencies fixed at equality. The mean and variance of the estimates were compared to the prediction from the asymptotic normal distribution.

Secondly we studied the estimate of the transition/transversion rate ratio under Kimura's 2 parameter model. Simulations were performed at various divergences and the parameter modelling the transition/transversion rate ratio and the divergence were both estimated. Again the mean and variance of the estimates were compared to the prediction from the asymptotic normal distribution.

Next we note that when studying the non-synonymous/synonymous rate ratio, setting ω_0 to 1 gives us the null hypothesis that the rate of synonymous changes equals the rate of non-synonymous ones. Equivalently, when studying the transition/transversion rate ratio, setting κ_0 to 1 gives us the null hypothesis that the rate of transitions equals the rate of transversions. These are both special cases of the more general models that allow the rate ratios to vary. Thus we use equation 3.7 to investigate the probability of both not rejecting the null hypothesis when it is false (the type II error) at large divergences between two sequences and the probability of rejecting the null hypothesis when it is true (the type I error) at large divergences. Again we simulate under a true, known model at a true, known divergence. From above we expect the type I error rate to remain at 0.05, but cannot predict the type II error rate.

This standard theory cannot be directly applied to the tree estimation problem since different trees define different non-nested models (YANG *et al.* 1995a). To investigate the effect of divergence we simulate 4-sequence trees using the subset of tree shapes (YANG 1998b) shown in Table 3.1 and calculate the expected probability of reconstructing the correct tree. The models used for the tree simulation were the HKY+γ model (HASEGAWA *et al.* 1985; YANG 1994b) and the codon model described above. The proportions of T, C, A and G nucleotides were 0.1, 0.2, 0.3, 0.4. Expected codon frequencies were calculated from these nucleotide frequencies. The transition/transversion rate ratio was set at 5.0. The phylogenetic reconstruction was done under HKY+γ in both cases.

The theory of (3.6) and (3.7) does not extend easily to methods based on the minimum change criterion. Therefore we simulate multiple 2-sequence data sets as above and calculated the mean and variance of the estimate of the non-synonymous/synonymous rate ratio using both the Nei-Gojobori method (NEI and GOJOBORI 1986) and the modified method by (ZHANG *et al.* 1998). We cannot compare the variance in the estimates to any expected variance. Also, instead of the likelihood ratio test shown in (3.7), we investigate the power of the method to reject the hypothesis of neutral evolution at high divergences using the Z-test as suggested in the original method and the Fisher Exact Test in the modified method.

When using counting methods to estimate the transition/transversion ratio, which we denote R^* to distinguish it from the transition/transversion rate ratio, the least variance unbiased estimator calculated from the count of changes is taken (INA 1998) as:

$$R^* = \frac{n\hat{p}}{n\hat{q}+1} \qquad \text{.....(3.8)}$$

Here $\hat{p}$ is the proportion of transitionally different sites and $\hat{q}$ the proportion of transversionally different sites. To test the power of this counting method to detect transition/transversion bias at high divergences in a statistically significant way, we use a Fisher Exact Test analogous to the test used for the nonsynonymous/synonymous rate ratio. For each site there are 2 possible transversions and one possible transition. Thus for a sequence of length m, there are $2m/3$ expected transversions and $m/3$ expected transitions. The observed number of transitional and transversional sites are compared to the expected numbers in the same way as the observed numbers of non-synonymous and synonymous sites are compared to the expected numbers in (ZHANG *et al.* 1998). This test is new and included here to complete the comparison between counting and likelihood methods. However it will be shown that the test performs surprisingly well.

Simulations were done using the program evolver from the PAML package. Maximum likelihood parameters were estimated using baseml and codeml from the same package. The Nei-Gojobori estimates was calculated using the program ng-new(ZHANG *et al.* 1998). The Fisher Tests were done using code modified from the R package available from http://www.ucl.ac.uk/~ucbpdcd/thesis/chapter3. Calculation of R^* was done with a simple perl script. Calculation of the expected information was done using the relation:

$$I_{jk} = E\left[\frac{\partial}{\partial\theta_j}\frac{\partial}{\partial\theta_k}\log P(x_i \mid \theta)\right]_{\theta^*} = -\sum_{x_i}\left[\frac{1}{P(x_i \mid \theta)}\frac{\partial}{\partial\theta_j}P(x_i \mid \theta)\frac{\partial}{\partial\theta_k}P(x_i \mid \theta)\right]_{\theta^*} \quad \ldots\ldots(3.9)$$

Numerical derivatives were calculated using the adaptive central difference algorithm of the GSL package, http://sources.redhat.com/gsl/. Four-sequence trees were calculated by calculating the maximum likelihood estimate for each of the 3 possible topologies with

baseml or calculating the parsimony score for each topology using pamp from the PAML package.

3.3 Results and Discussion

3.3.1 Saturation and the mean estimates

The difference in the mean estimate of the parameter of interest at high divergences (figure 3.1) provides evidence for two mechanisms of saturation (S.A.E. and S.C.V.). Since counting methods depend on minimising the number of changes, we would expect the mean estimate of the parameter of interest to be close to the true value at small branch lengths, but not necessarily at large branch lengths. In contrast, likelihood-based methods do not depend on this assumption and are theoretically asymptotically unbiased. Hence for large enough data sets we expect the mean likelihood-based estimate of the parameter of interest to be close to the true value even at large divergences. In our simulations this is indeed what happens: the expected value of the likelihood estimate remains close to the true value in all cases. However the expected value of the parsimony method estimate, whilst close to the true value at small divergences, deviates from the true value at large divergences in most of cases shown. The exception is the transition/transversion model with the transition/transversion rate ratio set to 1. Thus S.C.V. affects the expected value of the parsimony estimate but S.A.E. does not affect the expected likelihood estimate. We have two mechanisms of saturation and two qualitatively different behaviours.

The likelihood-based estimate is only asymptotically unbiased. The effect of small sequence length is shown in figure 3.2. There seems to be some bias in the maximum

likelihood estimate of the non-synonymous/synonymous rate ratio at both small and large divergences. Additionally there is some bias in the transition/transversion rate ratio at small divergences. However this bias is still smaller than the bias present in the counting method.

3.3.2 Saturation and the variance in the estimates

Whilst it seems that S.A.E. does not affect the expected value of the likelihood estimate, it does affect the variance as shown in figure 3.3. S.A.E. has a smaller effect on the variance of the likelihood estimate of the non-synonymous/synonymous rate ratio, than S.C.V. does on the variance of the counting method. However the parsimony based transition-transversion bias estimator (R^*) has a lower variance at high sequence divergence than its likelihood-based counterparts. In this case, we recall that the expected counting-based estimate deviates from the true value at high sequence divergence. This implies that the counting-based estimates are more tightly grouped around the wrong value, which is highly misleading.

The fit between the expected and observed variance in the likelihood estimate is shown in figure 3.4. For large sequences (n = 500 codons) there is a fit between the variance of the estimate of the non-synonymous rate ratio and the predictions from the asymptotic normal distribution. However for small sequences (n = 100 codons), the variance of the estimate is markedly higher than that predicted from the asymptotic result. The fit between the expected variance in the transition/transversion rate ratio estimate and the observed shows a similar pattern.

3.3.2 Saturation and hypothesis testing

The type I error rate for likelihood based models is more or less unaffected by S.A.E., as shown in figure 3.5a and 3.5b. The error rate remains at around 0.05, implying that the asymptotic approximations hold even at high divergences at the sequence lengths studied. However tests of significance based on the parsimony methods show an increase in the type I error rate at high sequence divergence. In the case of the transition/transversion bias estimate this increase is slight and possibly of little concern. In the case of the non-synonymous/synonymous rate ratio the increase is catastrophic. It is however consistent with the biased expected mean estimate shown in figure 3.3. The increase in type I error rate for the parsimony based method is alarming as it would lead one to reject the null hypothesis even when the data has a high probability of being generated under it. Hence one would erroneously infer selection. The two mechanisms of saturation produce two qualitatively different behaviours in the power of the test.

3.3.3 Saturation and phylogenetic reconstruction

The effect of S.C.V. and S.A.E. on tree reconstruction are shown in figures 3.6 and 3.7. It can be seen that in most cases, S.C.V and S.A.E. have more or less the same effect on tree reconstruction for parsimony and maximum-likelihood analyses. The exception is the "Felsenstein tree", the 3rd tree shape described, with long branches in different clades. Here S.C.V causes a rapid drop in the accuracy of the parsimony reconstruction.

One thing we note is the complex interplay between total tree length and the shape of the distribution describing the relative rate of evolution. When the shape parameter is

very small, a large number of sites are evolving slowly with a small number of very rapidly evolving sites. At low divergences the large number of slowly evolving sites have a low probability of having changed in a phylogenetically informative way. However at higher divergences these sites will have a higher probability of having changed, hence there will be a stronger phylogenetic signal from these sites. On the other hand the long tail of rapidly evolving sites will quickly approach their equilibrium probabilities, even at relatively small average divergences, lowering their phylogenetic signal. The complex interplay between these two effects accounts for the counter-intuitive shape of the probability of tree reconstruction at the divergences studied with highly skewed rate distributions, as confirmed in Figure 3.8. This shows the complex interplay at both vey high and very low divergences. As expected at very low divergences there is little phylogenetic signal and the probability of constructing the correct tree is approximately a third. At very high divergences the sequences become saturated and the probability of reconstructing the correct tree drops to about a third.

Finally, it is also worth noting that strong selection seems to have little effect on the probability of reconstructing the correct tree using a nucleotide model with variable rates.

3.4 Conclusion

The results provide some justification for the definition of two effects of saturation. Parsimony-based methods are prone to saturation by criterion violation because they rely on a minimum change criterion that is violated at large divergences. Thus as sequences diverge they show a decline in performance that can lead to the wrong conclusions being drawn from the data. This is shown by the bias in the expected parsimony-estimates and most importantly in the increase in the type I error rate. In contrast likelihood methods remain unbiased with predictable type I error rates. Likelihood methods are prone to S.A.E. which manifests itself as a mostly predictable increase in the variance of the estimate and a reduction in the power of the tests. When saturation affects likelihood methods, it does so predictably. It is perhaps unsurprising that parsimony-based methods fail when the fundamental criterion on which they are based is violated. Whilst it may seem unreasonable to expect parsimony-based methods to work under conditions for which they are not designed, we are inevitably sometimes faced with the question of how to analyse highly divergent sequences. It seems that likelihood methods are a better choice for these situations because one is unlikely to reject null hypotheses without adequate evidence.

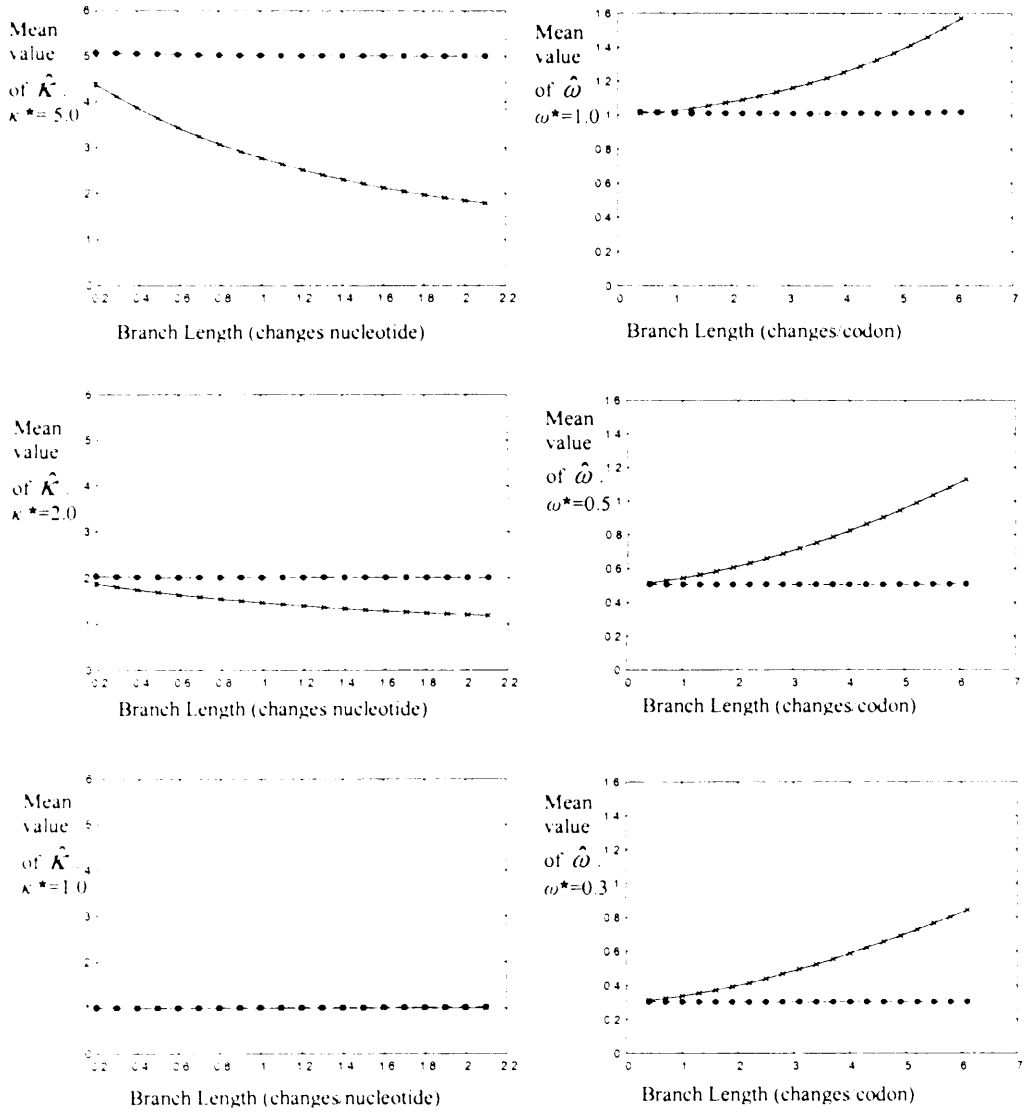


Figure 3.1. Observed mean value of the estimate of the transition/transversion rate ratio (κ, left) and non-synonymous/synonymous rate ratio (ω, right) for two sequences. The true values, κ^* and ω^* are shown. Values for the likelihood estimator are shown by dots and dotted lines and values for the counting method by crosses on solid lines. Sequence length is 1500 nucleotides for κ and 500 codons for ω.

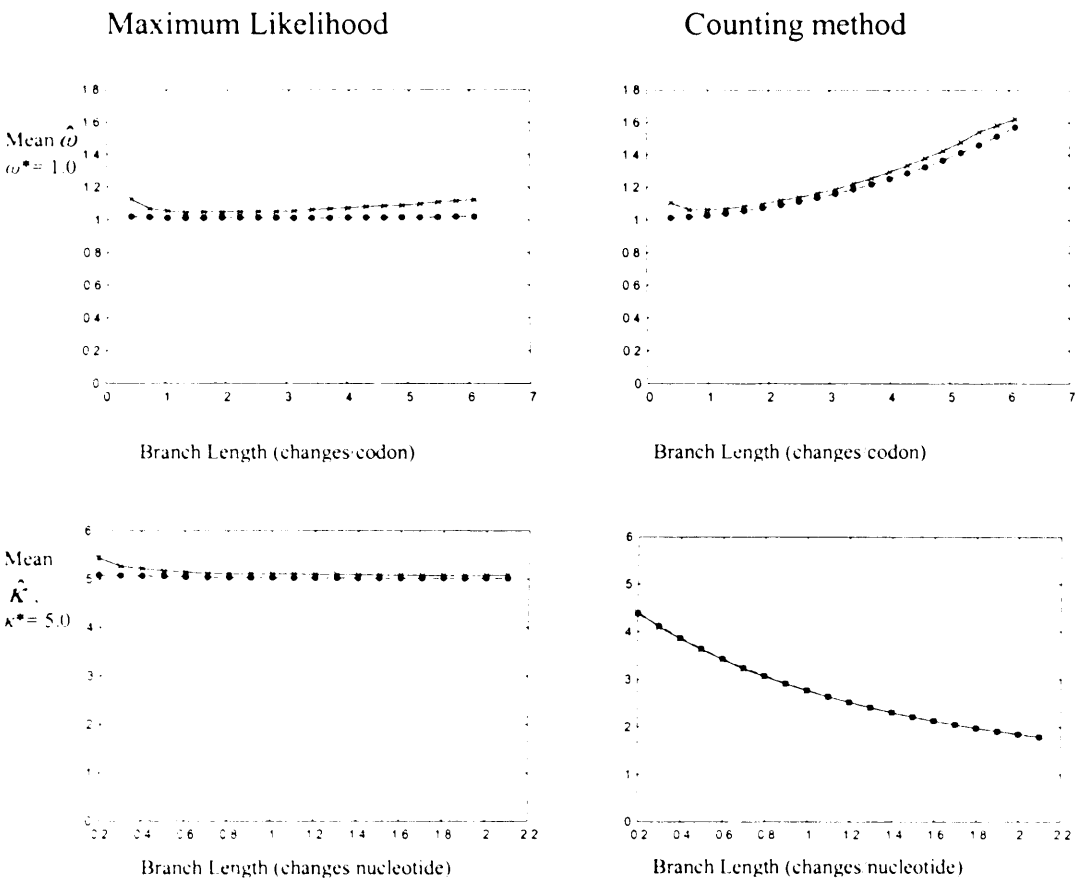


Figure 3.2. The mean of the estimates of the transition/transversion rate ratio (κ, bottom) and non-synonymous/synonymous rate ratio (ω, top) for two sequences plotted against the branch lengths. The true values, κ^* and ω^* are shown. Mean estimates for a sequence length of 1500 nucleotides (500 codons) are shown by dotted lines and for a sequence length of 300 nucleotides (100 codons) by solid lines. Estimates calculated by maximum likelihood methods are shown on the left, those calculated by the counting methods on the right.

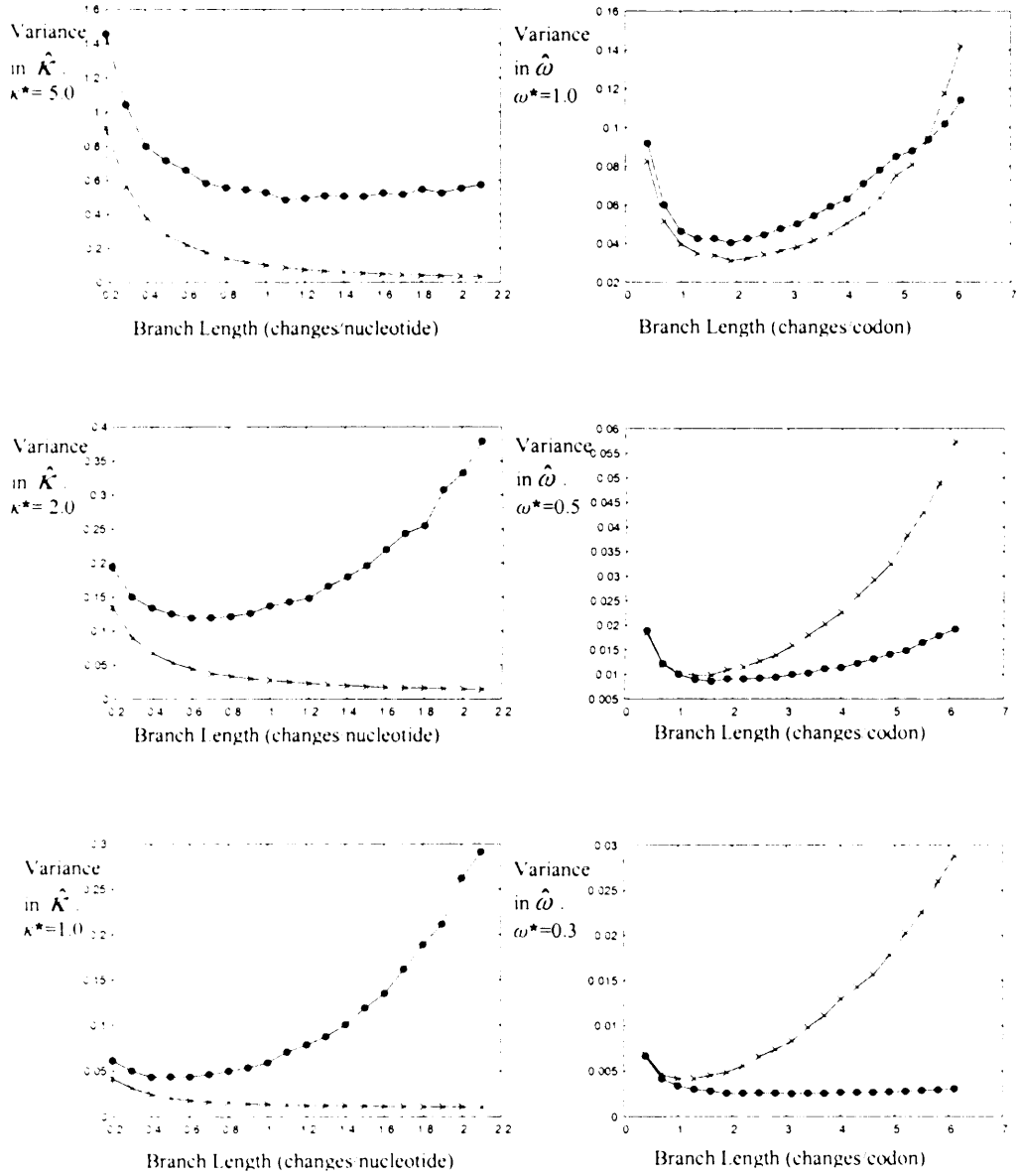


Figure 3.3. Observed variance in the estimate of the transition/transversion rate ratio (κ, left) and non-synonymous/synonymous rate ratio (ω, right) for two sequences. Variances for the parsimony estimator are shown by crosses and dotted lines, likelihood by dots and solid lines. The sequence length is 600 nucleotides (200 codons).

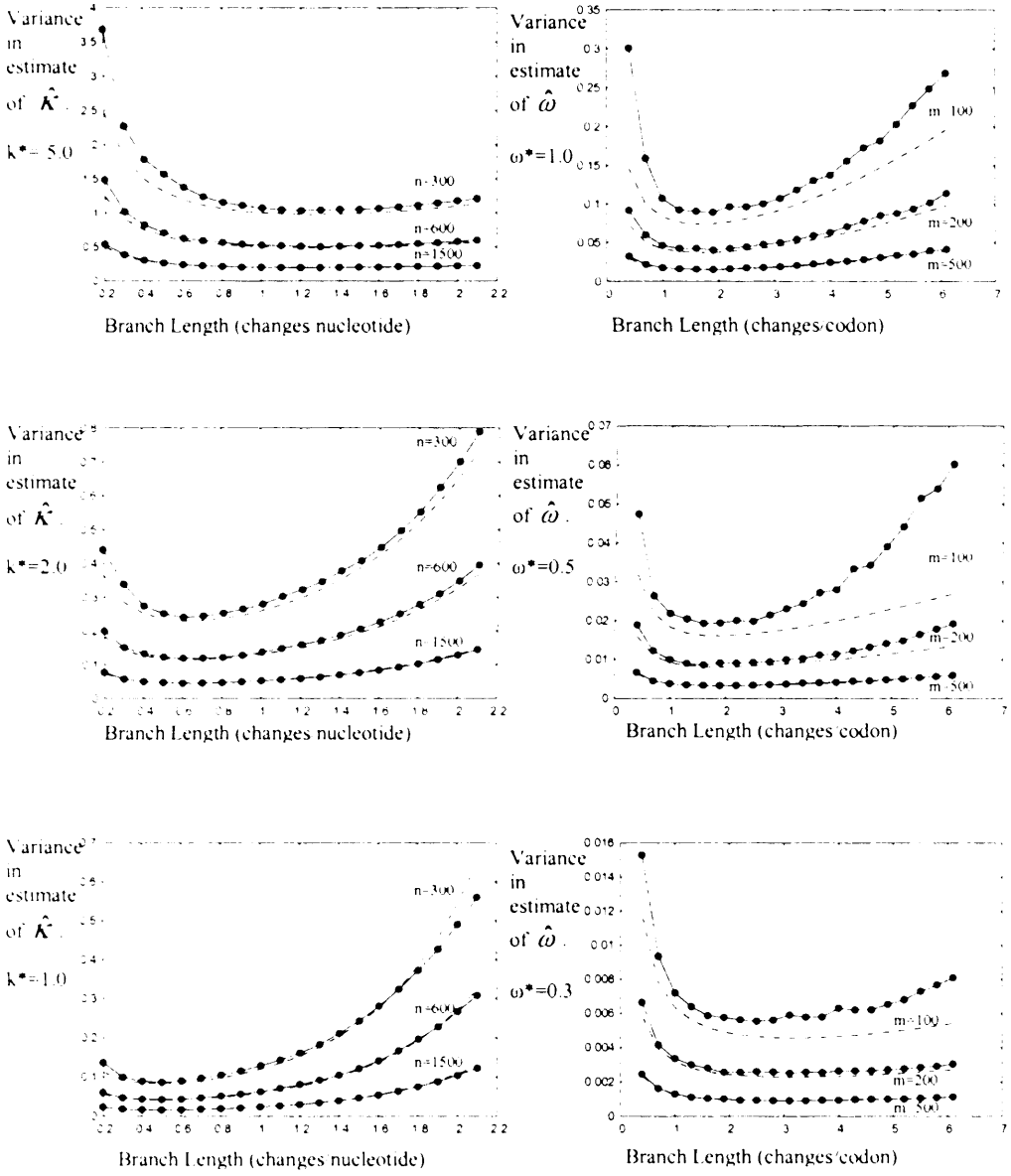


Figure 3.4. Observed (solid lines) and expected (dotted lines) variances of the transition-transversion rate ratio (left) and the non-synonymous/synonymous rate ratio (right). Sequence length *n* is the number of nucleotides nucleotides, *m* is the number of codons.

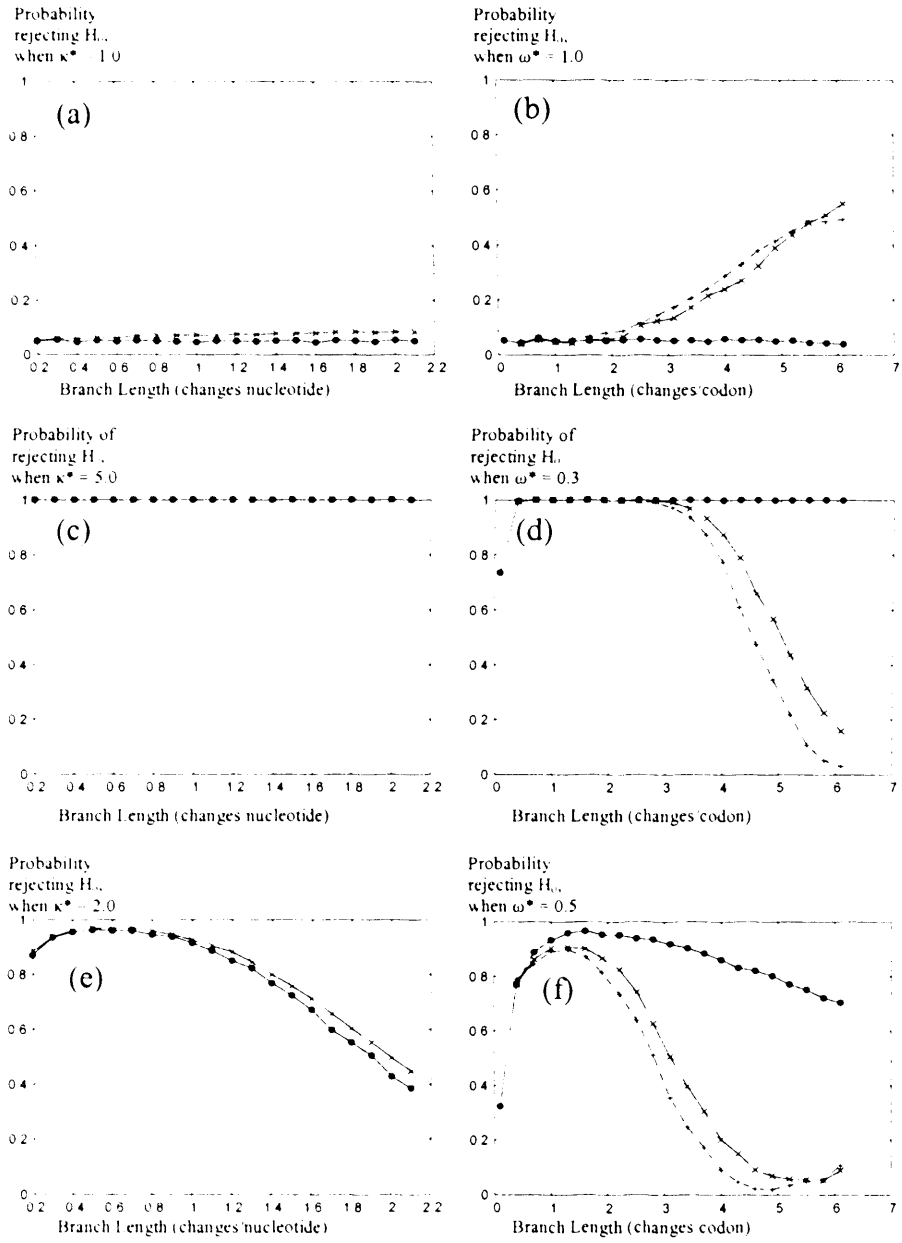


Figure 3.5. Observed probability of rejecting the null hypothesis of unbiased transition/transversion rate ratio (left) and non-synonymous/synonymous rate ratio equal to 1 (right). Sequence length is 600 nucleotides (200 codons). Results for likelihood shown by solid lines and dots, counting method by dashed lines and diagional crosses. Results for Nei-Gojobori method shown by dotted lines and vertical crosses on the right hand set of graphs.

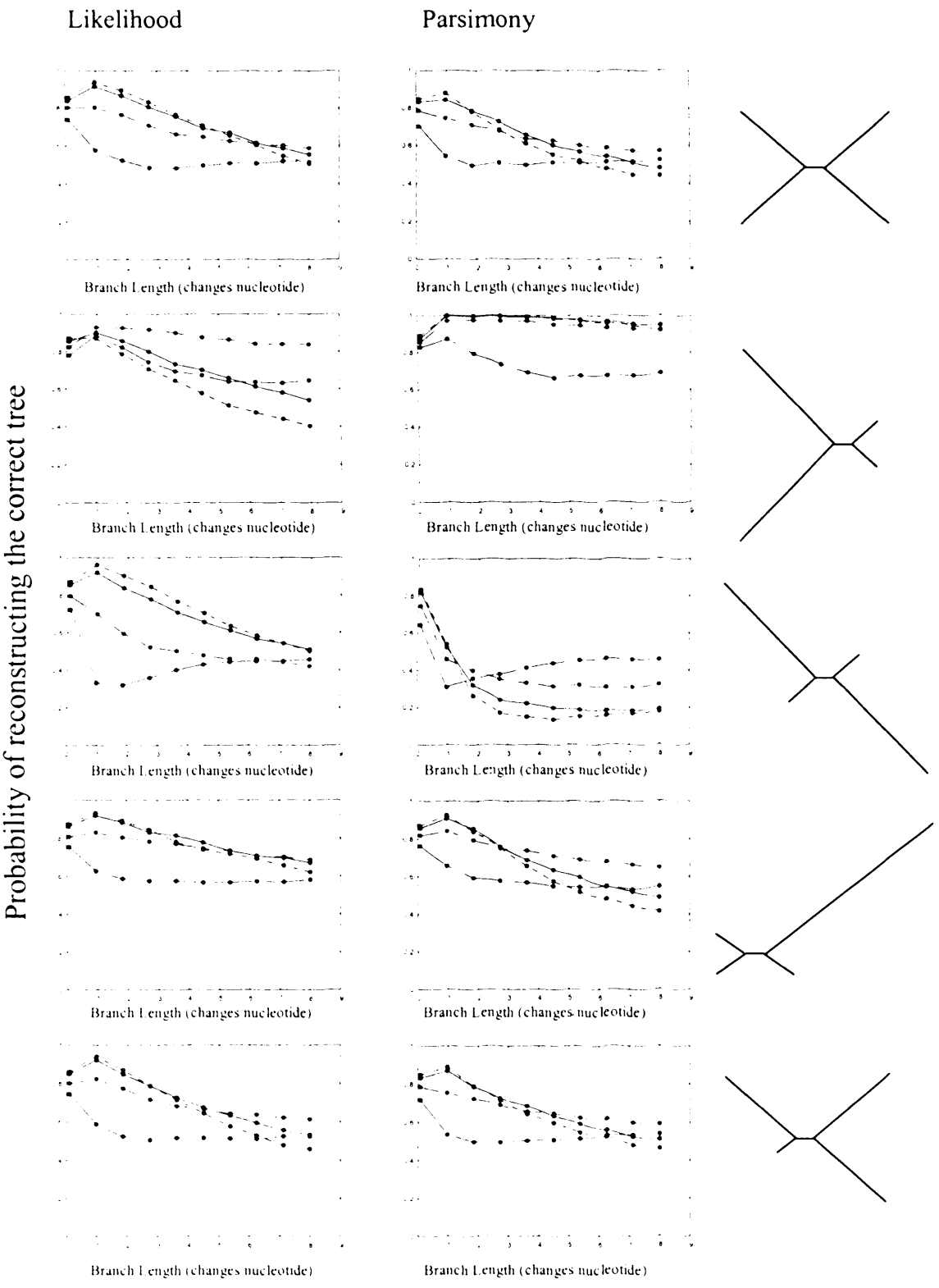

Likelihood
Parsimony
Probability of reconstructing the correct tree
Branch Length (changes nucleotide)
Branch Length (changes nucleotide)
Branch Length (changes nucleotide)
Branch Length (changes nucleotide)
Branch Length (changes nucleotide)
Branch Length (changes nucleotide)
Branch Length (changes nucleotide)
Branch Length (changes nucleotide)
Branch Length (changes nucleotide)
Branch Length (changes nucleotide)

Figure 3.6. Probability of reconstructing the correct tree using likelihood (left) and parsimony methods (right). Tree length given on x-axis is in expected number of nucleotide changes per site. Both the simulation model and model used for likelihood reconstruction is HKY85 + γ. Shape parameter α=10 indicated by dotted lines, α=0.464 by small dashes, α=2.15 by solid lines, α=0.1 by large dashes. Tree shapes shown.

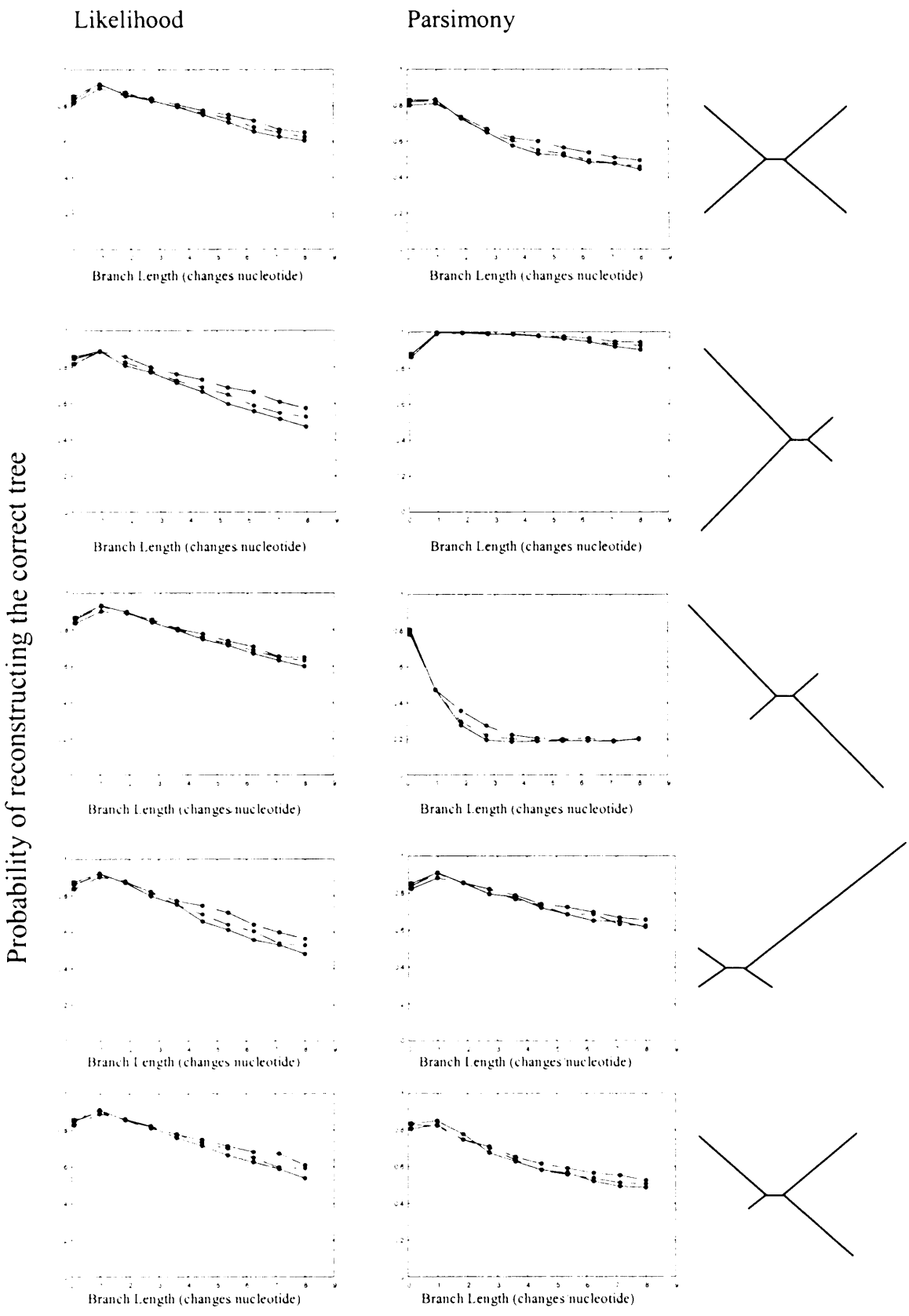

Likelihood
Parsimony
Probability of reconstructing the correct tree
Branch Length (changes nucleotide)
Branch Length (changes nucleotide)
Branch Length (changes nucleotide)
Branch Length (changes nucleotide)
Branch Length (changes nucleotide)
Branch Length (changes nucleotide)
Branch Length (changes nucleotide)
Branch Length (changes nucleotide)
Branch Length (changes nucleotide)
Branch Length (changes nucleotide)

Figure 3.7. Probability of reconstructing the correct tree using likelihood (left) and parsimony methods (right). Tree length given on x-axis is in expected number of nucleotide changes per site. True model is a codon model with ω=1.0 indicated by solid lines, ω=0.5 by small dashes and ω=0.3 by large dashes. Tree shapes shown. Model used for likelihood reconstruction is HKY + γ.

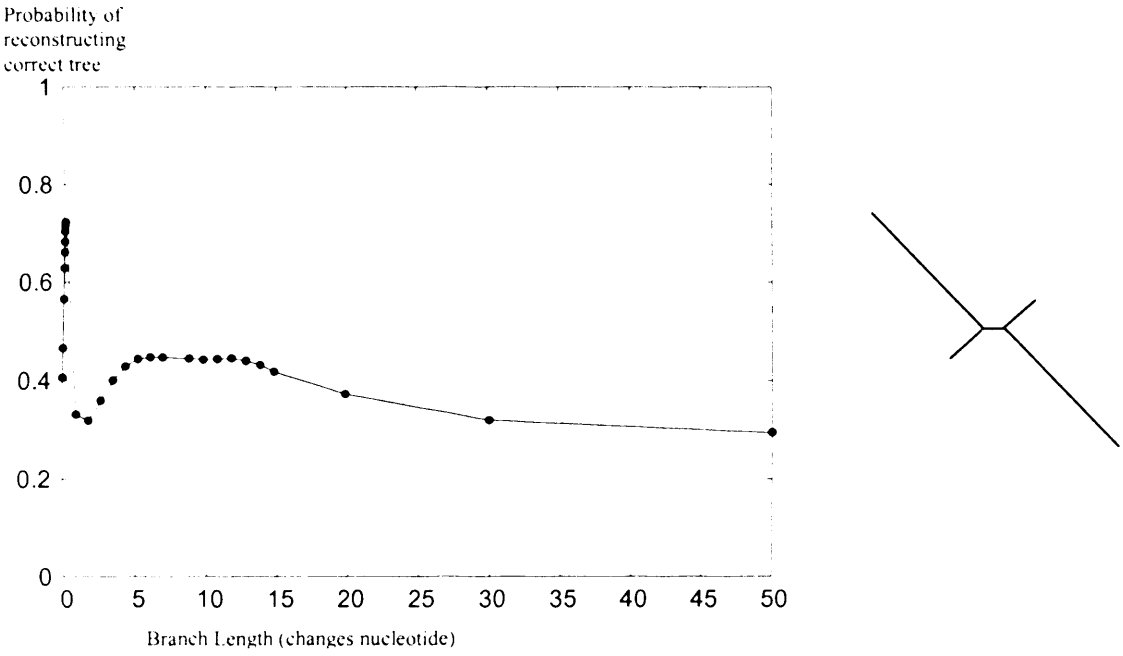


Figure 3.8. Probability of reconstructing the correct tree using likelihood. Tree given on x-axis in expected number of nucleotide changes per site. True model and model used for likelihood reconstruction is HKY85+γ. Skew parameter $\alpha = 0.1$. Tree shape shown.

	Internal	External A	External B	External C	External D
Tree 1	0.05 x L	0.2375 x L	0.2375 x L	0.2375 x L	0.2375 x L
Tree 2	0.05 x L	0.1 x L	0.1 x L	0.375 x L	0.375 x L
Tree 3	0.05 x L	0.1 x L	0.375 x L	0.1 x L	0.375 x L
Tree 4	0.05 x L	0.1 x L	0.1 x L	0.1 x L	0.6 x L
Tree 5	0.05 x L	0.3 x L	0.3 x L	0.3 x L	0.05 x L

Table 3.1. Shows the relative lengths of branches of trees studied. Each tree has topology {{A,B}, C, D}. The branch labelled External A is the branch leading to node A, the branch labelled External B leads to node B etc. Thus the long-branch attraction tree is Tree 3. The total tree length is marked as L.

CHAPTER 4. THE EFFECT OF HETEROTACHY ON TREE RECONSTRUCTION BY PARSIMONY AND LIKELIHOOD

4.1 What is heterotachy?

A site in a molecular sequence is said to be undergoing heterotachous evolution if its evolution is determined by a process whose rate varies in time relative to the other sites in that sequence. For example consider a site that switches between a mutable state where changes are possible and a non-mutable state where changes are not. If the other sites in the sequence remain mutable, the original site is undergoing heterotachous evolution (GALTIER 2001; HUELSENBECK 2002) since its rate is varying relative to that of the other sites. Alternatively if a single site is rapidly evolving in one part of a tree but slowly evolving in another whilst other sites evolve at the same rate across the tree the site is said to be undergoing heterotachous evolution (PUPKO and GALTIER 2002). Heterotachous evolution of single sites is set in contrast to between-site rate heterogeneity models such as the gamma model (YANG 1994b) and the rate auto-correlation model (FELSENSTEIN and CHURCHILL 1996; YANG 1995). These allow different sites on a sequence to be probabilistically assigned to classes of sites, each class evolving at a different rate. However in these models, the rate for a class of sites does not change over time relative to the other classes. In other words, slowly evolving sites stay slowly evolving and rapidly evolving sites stay rapidly evolving.

Heterotachy has been detected by testing whether the pattern of substitutions observed in different subtrees of a phylogeny fit the pattern expected under the non-heterotachous gamma rate heterogeneity model (LOCKHART *et al.* 1998; MIYAMOTO and

FITCH 1995). More recently a method was developed that tests whether substitutions appear dispersed evenly across a tree (LOPEZ *et al.* 2002). Heterotachy has also been modelled directly using a likelihood model with a parameter to model the rate of switching between mutable and non-mutable states (GALTIER 2001; HUELSENBECK 2002). Explicit parametric modelling allows the use of the likelihood ratio test in a standard statistical approach.

Heterotachy has been of interest because it has been shown (INAGAKI *et al.* 2004) that removing heterotachous sites affects phylogeny reconstruction when the reconstruction is performed using the non-heterotachous HKY85 model (HASEGAWA *et al.* 1985). It had been mostly believed (SULLIVAN and SWOFFORD 2001) that likelihood models were consistent under mild model violations.

4.2 Heterotachy and phylogeny reconstruction

4.2.1 Heterotachy: the argument for parsimony

Recently Kolacskowski and Thornton (KOLACZKOWSKI and THORNTON 2004) showed that under extreme conditions of heterotachy, non-heterotachous likelihood models were inconsistent, giving strong support for the wrong tree. In the situation they described (see Figure 4.1), the sites were divided into two classes that evolved according to two different sets of branch lengths. Under these conditions, they argued that parsimony-based phylogeny reconstruction performed better than non-heterotachous likelihood-based reconstruction. Additionally they made some attempt to develop a weighted likelihood model to account for heterotachy. In this model, the probability of observing the data given that it belonged to a particular heterotachous class, was weighted by the posterior

probability of the data belonging to that class. They found that this non-standard likelihood model still under-performed parsimony reconstruction methods and was inconsistent even when they believed it correctly described the evolutionary process. This was regarded as of interest since all correctly specified likelihood models can be shown to be consistent, see for example (CRAMER 1946).

4.2.2 Likelihood and heterotachy

It has since been shown (SPENCER *et al.* 2005) that the conditions of the original study were such that a general conclusion that parsimony out-performed likelihood in heterotachous situations could not be drawn for a number of reasons. Firstly, the evolutionary model used in the original study to both simulate the data and draw inferences from it was that of Jukes-Cantor (JUKES 1969). This is the simplest, least biologically realistic model of nucleotide evolution. It has been argued in simulation studies (YANG 1996) that parsimony reconstruction is close to performing likelihood reconstruction with the Jukes-Cantor model. Hence simulating sequences under this model is likely to give parsimony an unrealistic advantage that it would not receive from real biological sequences. Secondly Spencer et al. showed that the tree shapes chosen to demonstrate heterotachy were exceptional in that they caused the likelihood method to fail without affecting the parsimony reconstruction. With other tree shapes likelihood outperformed parsimony, even though the data was generated under heterotachous conditions. Finally the weighted-likelihood model of the original paper was shown by Spencer et al. to be outperformed by a correctly specified mixture model. These results were confirmed in (GADAGKAR and KUMAR 2005). We defer the discussion of the exact form of the mixture

model to the methods section as it is similar to the model used in this chapter. The standard mixture model of Spencer et al. did not suffer from the inconsistency problem that the incorrect weighted-likelihood model does.

In this study, which predates the work of Spencer et al., we address the effect of heterotachy on phylogeny reconstruction by parsimony and likelihood methods. We focus here on the four-sequence case, as shown in Figure 4.1. It has already been shown that parsimony-based methods can be misleading when sequences undergo rapid evolution along single terminal branches that are in different clades. When the reconstruction is performed, the most parsimonious tree incorrectly places both long branches in the same clade. Indeed if the long branches are sufficiently long, parsimony becomes inconsistent, i.e. as more data is added it becomes more and more likely that parsimony will reconstruct the wrong tree. Likelihood methods are not so vulnerable to this artefact. This is the well-known long-branch attraction problem (FELSENSTEIN 1978). In contrast, it has been shown (CHANG 1996) that a likelihood method that incorrectly specifies rate variation can be inconsistent as well. Thus in the case originally studied by Kolackskowski and Thornton, each class of sites undergoes rapid evolution along a different branch in the different clades but the likelihood models do not describe the rate variation corectly. Thus in this case we have conditions that cause problems for both likelihood and parsimony methods.

4.3 The biological significance of heterotachy

Heterotachy is of broader interest than solely its effect on phylogeny reconstruction. It has been argued (MIYAMOTO and FITCH 1995) that shifts in a site's rate of evolution reflect shifts in the selective regime under which that site is evolving. For example a nucleotide site that forms part of a codon coding for an amino acid essential for the structure of a protein may be under strong purifying. If that selection is relaxed, the site will be free to accumulate mutations more frequently and will undergo a heterotachous increase in its rate of evolution. It is worth noting that heterotachy per se does not necessarily imply a functional shift (PHILIPPE *et al.* 2003) since broader genomic and population-level processes may be at work. For example different regions of a chromosome evolve at different rates, see for example (DE BAERE *et al.* 2003), and a heterotachous shift in a sequence's evolution may simply reflect a chromosomal rearrangement. However, this specific point aside, it seems that the ability to assign different sites within a sequence to different heterotachous classes would allow some inferences about the selective regime to be drawn if coupled with other biological knowledge. In this study, as well as investigating phylogeny reconstruction, we investigate the power of the mixture model to (1) correctly detect whether any sites in a sequence have undergone heterotachous evolution and (2) to correctly detect specifically which sites have under gone heterotachous evolution.

4.3 Methods

4.3.1 The mixture model of heterotachy

In this section we introduce a mixture model of heterotachy that is broadly similar to that of Spencer et al.(SPENCER *et al.* 2005) but with extra parameters that more realistically model DNA sequence evolution. The instantaneous rate of transition between states on a branch is described by the HKY85 model. This has parameters describing both the nucleotide frequencies and the transition/transversion rate ratio. Between site rate-heterogeneity was described using a parameter determining the shape of a gamma distribution with mean 1.0 as described by (YANG 1994b). Heterotachy was modelled by including two classes of branch lengths [see Figure 4.1], and a single parameter (ρ) to describe the probability that a site belongs to one of the classes. The probability that a site belongs to the other class is given by ($1-\rho$).

In the model the tree topology is assumed to be the same for each branch class. Each branch class describes a different set of evolutionary distances between the nodes in the topology. The model can be represented:

$$P(x_i;\theta,\alpha) = P(x_i \mid t^0;\theta,\alpha)\rho + P(x_i \mid t^1;\theta,\alpha)(1-\rho) \quad \text{.....(4.1)}$$

Here x_i is the data observed at the tree tips at a site, θ represents the parameters of the instantaneous rate matrix, t^0 the branch lengths of the first class, t^1 the branch lengths of the other and α the shape parameter of the Gamma distribution that models between-site rate heterogeneity. To estimate the values of these parameters we use the method of

maximum likelihood, i.e. we choose the values of the parameters that maximise the probability of observing the data.

For an unrooted tree of m sequences, the non-heterotachous HKY + γ model has $2m$-3 branch length parameters, 4 parameters in the evolutionary rate matrix, and 1 parameter describing between-site rate heterogeneity. The new model has $4m$-6 branch length parameters, 4 parameters in the rate matrix, 1 parameter describing between-site rate heterogeneity and 1 parameter describing the probability of belonging to a particular class of branch lengths. It can be seen that the non-heterotachous model is nested within the new model, having an extra $2m$-2 parameters. Thus if the non-heterotachous model is our null hypothesis we can test whether the null hypothesis adequately explains the data using a χ^2 distribution with $2m$-2 degrees of freedom.

4.3.2 Simulation to examine method performance

To test the efficacy of this model on phylogeny estimation we undertook a simulation using software available from http://www.ucl.ac.uk/~ucbpdcd/chapter4/. The object of the simulation was to estimate the probability of successfully reconstructing the true tree using different methods. Data were simulated on a 4 sequence tree with two classes of branch lengths. The sequence length was 5000 base pairs, with half the sites evolving on one set of branch lengths, half the sites on the other. Firstly we used the heterotachous model with Jukes-Cantor parameters to both simulate the data and reconstruct the phylogeny. Next we used the more realistic heterotachous version of the HKY 85 + γ model (Equation 4.1) to both simulate the data and reconstruct the phylogeny. When reconstructing the phylogeny, the probability of observing the data was calculated

for each of the 3 possible topologies. If the true tree, i.e. the one on which the data was simulated, had the highest likelihood then the count of successes was incremented by one. If the true tree and one other topology both had the joint highest likelihoods the count of successes was incremented by a half and if all three topologies had the same likelihood the count of successes was incremented by a third. The proportion of correct topology reconstructions was then the count of successes divided by the total number of trials and this was taken as the estimate of the probability of successfully reconstructing the true tree. Additionally similar analyses were performed using the heterotachous Jukes-Cantor model for simulating the data but using the non-heterotachous Jukes-Cantor model for tree reconstruction.

To test the consistency of the standard likelihood model we calculated:

$$\sum_i p^2_{T_0}(x_i)\log[\hat{p}^1_{T_0}(x_i)] - \sum_j p^2_{T_0}(x_j)\log[\hat{p}^1_{T_1}(x_j)] = \sum_i p^2_{T_0}(x_i)\log\left[\frac{\hat{p}^1_{T_0}(x_i)}{\hat{p}^1_{T_1}(x_i)}\right] \quad \text{.....(4.2)}$$

Here T_0 is the true topology and ${p_{T_0}}^k(x_i)$ is the probability of observing the data, x_i, under a model with k classes of branch lengths on T_0 with parameters set at their true value. When k=1, the model has 1 class of branch lengths, ie it is not heterotachous. When k=2 the model is the heterotachous model described in equation 4.1. ${\hat{p}_{T_n}}^k(x_i)$ is the probability of observing the data x_i under the k class model, on topology T_n, given that all the parameters in the model have been estimated at their "equilibrium maximum likelihood value". We define the equilibrium maximum likelihood value as the parameter value that maximises the likelihood of the data when the data is comprised of every possible datum, each

observed at a frequency equal to the probability of observing it under the true model. Thus $p_{T_0}{}^{k}(x_i) = \hat{p}_{T_0}{}^{k}(x_i)$ because likelihood is consistent when the model is correct. It can be seen that this measure is the expected difference in likelihood between the true tree and an incorrect tree, and is related to the Kulback-Leibler measure of distance between two probability distributions.

The tree topologies initially tested are shown in figure 4.1. In addition we performed a random branch simulation. Again we used a four-sequence topology. However the external branch lengths were drawn from a Uniform(0,5] distribution with the internal branch length fixed.

To test the power of the model to detect heterotachous evolution in a realistic scenario we looked at trees of 30 sequences. The 43 topologies of 30 sequences were selected from release version 2 of the Pandit database (WHELAN *et al.* 2003). For each topology we sampled branch lengths from a Uniform(0,3] distribution and assigned them to two branch length classes. We simulated 500ρ sites along one class of branch lengths and 500 $(1-\rho)$ sites along another using the HKY85 + γ model. Here p is a parameter between 0 and 0.5. The sequence length of 500 was chosen because the mean sequence length in PANDIT is 453 sites. We then performed a maximum likelihood calculation using the true topology under the HKY + γ non-heterotachous likelihood model to determine the probability of the data under the null hypothesis of no heterotachy. This was performed using the PAML program (YANG 1997). We then performed a likelihood calculation using the mixture model. Since the models are nested, we used the fact that twice the log likelihood difference between the null and heterotachous model is asymptotically $\chi 2$

distributed with $2m$-2 degrees of freedom to test whether the null hypothesis of no evidence for heterotachy could be rejected.

To attempt to find specific sites belonging to particular heterotachous classes we performed the following calculation:

$$P(c[t^{j}]|x_i) = \frac{P(x_i|c[t^{j}])P(c[t^{j}])}{\sum_k P(x_i|c[t^{k}])P(c[t^{k}])} \quad \text{.....(4.3)}$$

Here $c[t^j]$ is the class with branch lengths t^j. $P(c[t^j])$ is the probability that a site belongs to the class with branch lengths t^j, i.e. $P(c[t^j]) = \rho$ if $j = 0$ and $P(c[t^j]) = 1-\rho$ if $j=1$, in the notation of Equation 4.1. $P(x_i|\ c[t^j])$ is the probability of observing the data at site i given that it belongs to the heterotachy class j with branch lengths t_j. Equation 4.3 was used in the 30 sequence simulation described previously whenever the likelihood ratio test detected heterotachy and sites were allotted to classes.

4.4 Results and Discussion

Under the extreme heterotachous conditions of figure 4.1, we see from figure 4.2 that when using the heterotachous Jukes-Cantor model for simulation, parsimony outperforms the non-heterotachous Jukes-Cantor model. Indeed it can be seen that the non-heterotachous likelihood model is inconsistent under these conditions from the expected log-likelihood difference between the true tree and the most likely wrong tree (figure 4.4). However the mixture model, which here is the same as that of Spencer et al., markedly outperforms both parsimony and the non-heterotachous likelihood model. When the more

realistic HKY85 + γ model is used for simulation and tree reconstruction, as shown in figure 4.3, parsimony still outperforms non-heterotachous likelihood but to a lesser extent. Again the heterotachous model performs best.

When the random branch simulation is performed, the heterotachous model does slightly less well than both the non-heterotachous model and parsimony (figure 4.5). Under these conditions once again non-heterotachous likelihood out performs parsimony. The random branch simulation was repeated when the number of simulation classes (n=5) exceeded the number of model estimation classes (n=2). Here the performance of parsimony and the non-heterotachous likelihood model were approximately the same, though once again the heterotachous likelihood method did worse than either (figure 4.6).

It can be seen that the heterotachous model performs worse under the random branch simulation than under the conditions of extreme heterotachy described in the original study. This phenomenon is confirmed by figure 4.7, showing the expected likelihood difference between the true tree and the next most likely wrong tree. We would expect a large expected log likelihood difference between the true tree and the next most likely wrong tree to imply a high probability of reconstructing the correct tree. Initially, the expected likelihood difference under conditions of extreme heterotachy increases rapidly as the internal branch lengthens, but actually decreases as the internal branch gets longer still. This initial rapid increase occurs to a lesser extent with less extreme heterotachy, so that for small internal branch lengths the more heterotachous the true model, the greater the expected likelihood difference and hence the greater the probability of reconstructing the true tree. Since a randomly generated tree is unlikely to be as heterotachous as those in the

original study of Kolackskowski and Thornton, this explains why the mixture model performs so well under those conditions.

The dip in expected likelihood difference for the heterotachous case was unexpected since when both the simulation and the estimation models are non-heterotachous, increasing the internal branch length allows the tree to be reconstructed with greater certainty. To confirm the existence of the dip we ran a set of simulations at varying degrees of heterotachy, but with very short sequence lengths. The short sequence lengths were necessary in order to ensure that at the values of internal branch length at which the dip occurred the probability of reconstructing the correct tree was different enough from 1 for the reduction in the probability of reconstructing the correct tree to be noticeable. The results shown in figure 4.8 show that at certain internal branch lengths, increasing the internal branch length reduces the probability of constructing the correct tree. Also it shows that at those same branch lengths more heterotachous trees are easier to reconstruct than less heterotachous trees. This is what we expect from the expected log-likelihood difference calculation.

The power of the method to detect heterotachy was very high. When the proportion of sites in the first heterotachy class ranged between 0.1 and 0.5, the likelihood ratio test detected heterotachy in every case. However when the posterior probability of a site belonging to a class was calculated, a high number of errors were observed, (figure 4.9). This appears to be because the value of the maximum likelihood estimates were not close to the true values.

4.5 Conclusion

Our results show that the superior performance at phylogeny reconstruction of parsimony over non-heterotachous maximum likelihood models under the conditions of the original study (KOLACZKOWSKI and THORNTON 2004) was real. However the effect is peculiar to the particular situation studied. When a random selection of branch lengths are used to generate the data, the standard maximum-likelihood methods with no heterotachy in the model outperform parsimony in most situations. Thus it would seem that current non-heterotachous methodologies will perform well in most cases.

The heterotachous situation that causes the non-heterotachous likelihood models to become inconsistent is the situation in which the mixture model performs best. Thus in the specific cases where the heterotachous conditions of the original study are suspected it would seem reasonable to use the heterotachy model to evaluate a subset of those trees suggested by non-heterotachous likelihood models and parsimony methods. In cases where there is little heterotachy the cost of the extra parameters must be paid by a small decrease in the probability of reconstructing the correct tree. However whether this is offset by the large increase in the probability of reconstructing the correct tree in conditions where the heterotachous effect is large will depend on how rare those conditions are

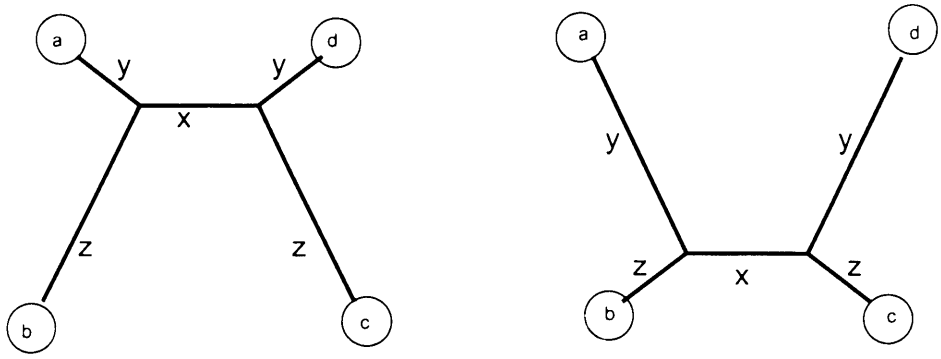


Figure 4.1. Heterotachous scenario of (KOLACZKOWSKI and THORNTON 2004). Each site is assigned to one of two heterotachy classes and allowed to evolve along one of the trees shown. In the simulations y was set at 0.05 changes/nucleotide, z at 0.75 changes/nucleotide.

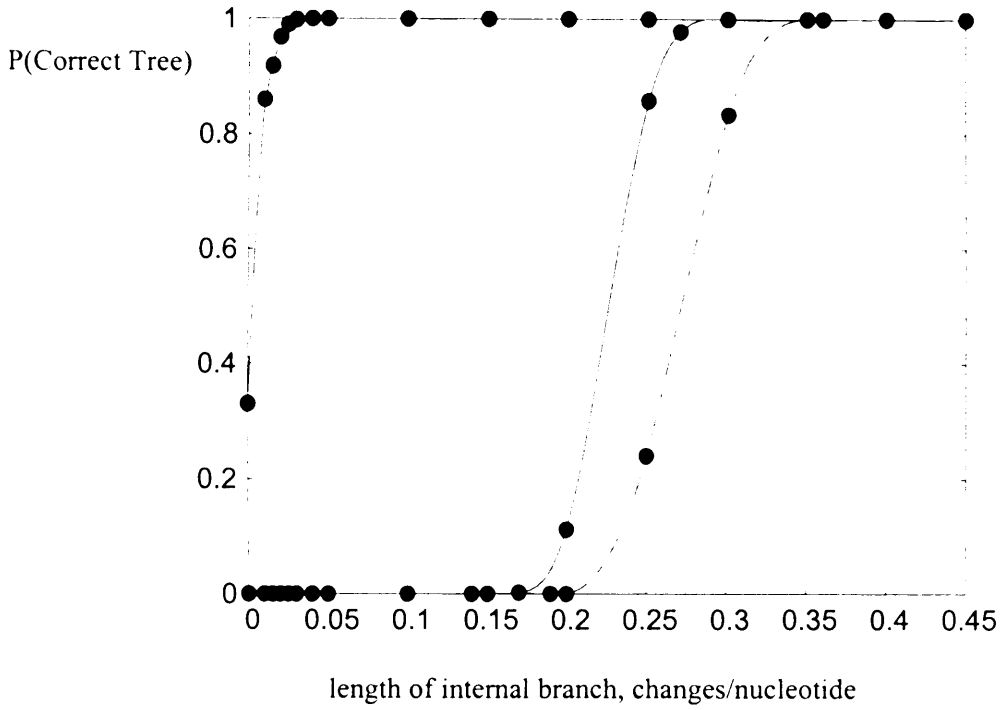


Figure 4.2. Probability of reconstructing the correct tree plotted against the length of the internal branch when the true model follows the heterotachous pattern shown in Figure 4.1. Parsimony shown by solid lines, non-heterotachous likelihood shown by dotted lines, heterotachous likelihood by dashed lines. The Jukes-Cantor model was used for both the simulation and estimation models.

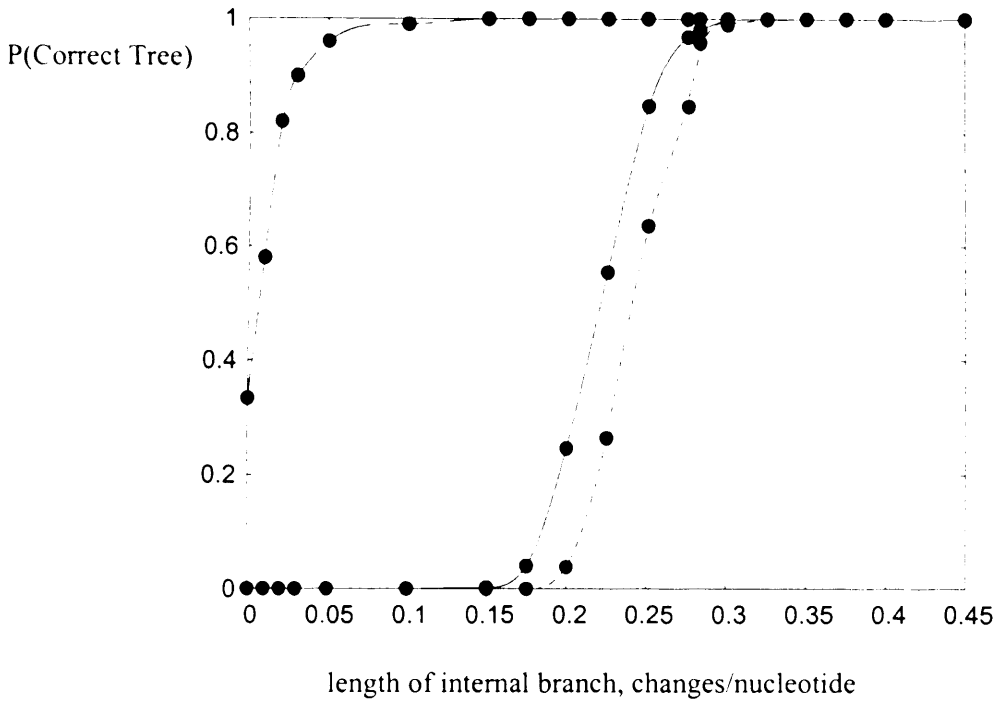


Figure 4.3. Probability of reconstructing the correct tree plotted against the length of the internal branch when the true model follows the heterotachous pattern shown in Figure 4.1. Parsimony shown by solid lines, non-heterotachous likelihood shown by dotted lines, heterotachous likelihood by dashed lines. The HKY85 + γ model was used for both the simulation and estimation models.

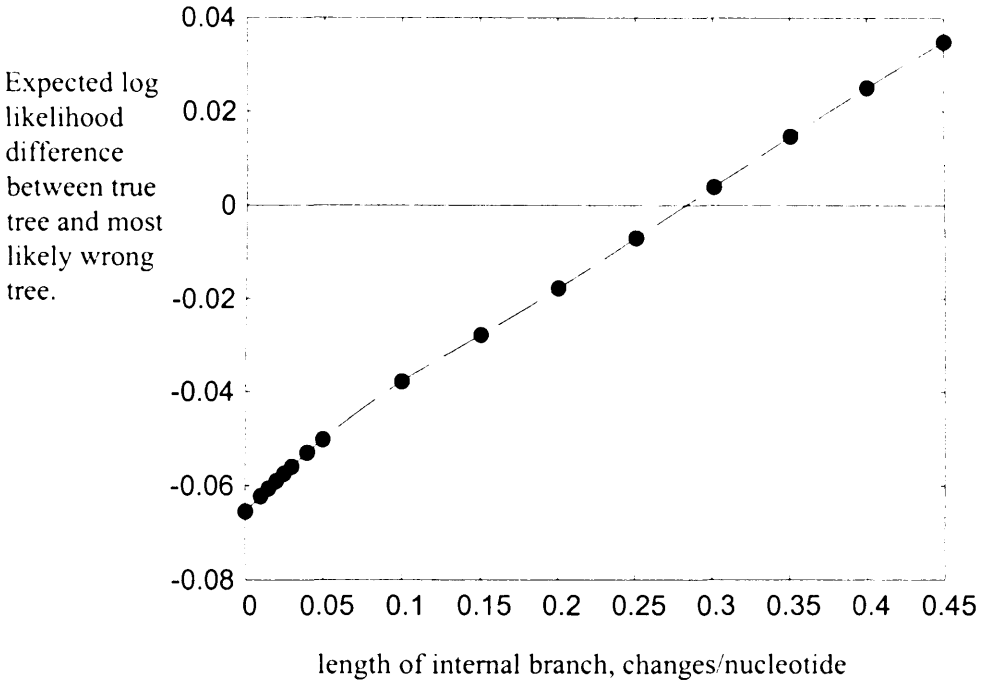


Figure 4.4. Expected log likelihood difference between the true tree and the most likely wrong tree when the true model follows the heterotachous conditions of Figure 1 but the estimation model has only one class of branch lengths. The Jukes-Cantor model was used for calculating both the probability of observing the data under the true tree topology and true heterotachous model and the probability inferred under the alternative topologies and non-heterotachous model.

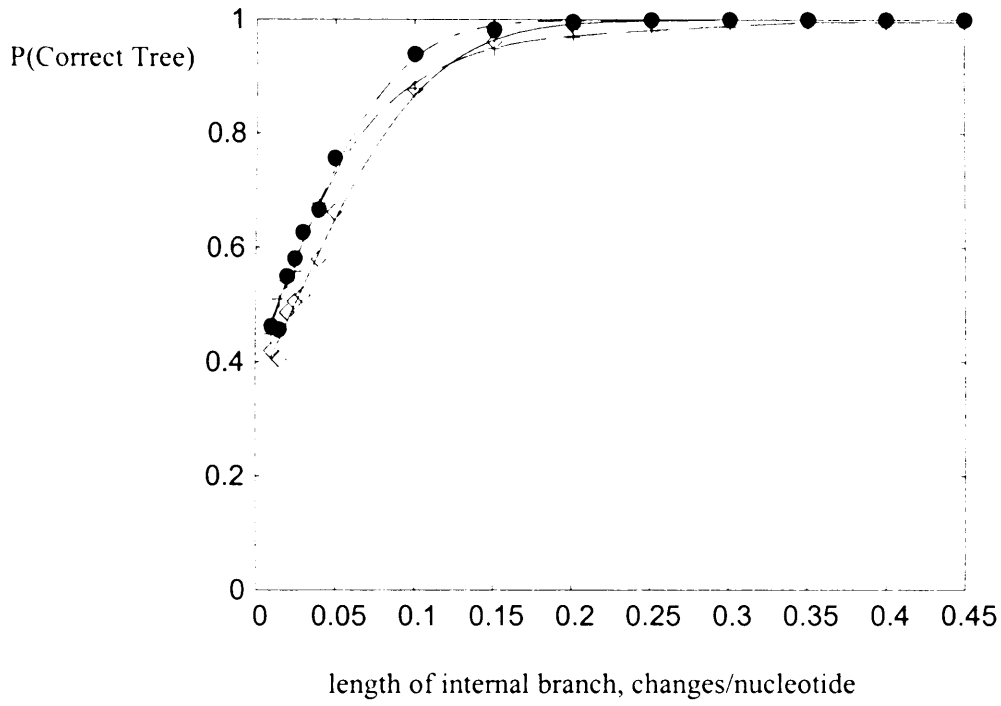


Figure 4.5. Probability of reconstructing the correct tree plotted against the length of internal branch under a 2-class random branch simulation. The simulation conditions were heterotachous and similar to those of Figure 2 but external branch lengths were drawn from a uniform distribution: U(0, 5]. Phylogeny reconstruction was done using parsimony (solid line), heterotachous likelihood (dashed line) and traditional non-heterotachous likelihood (dotted line). The Jukes-Cantor model was used for both simulation and likelihood-based reconstruction.

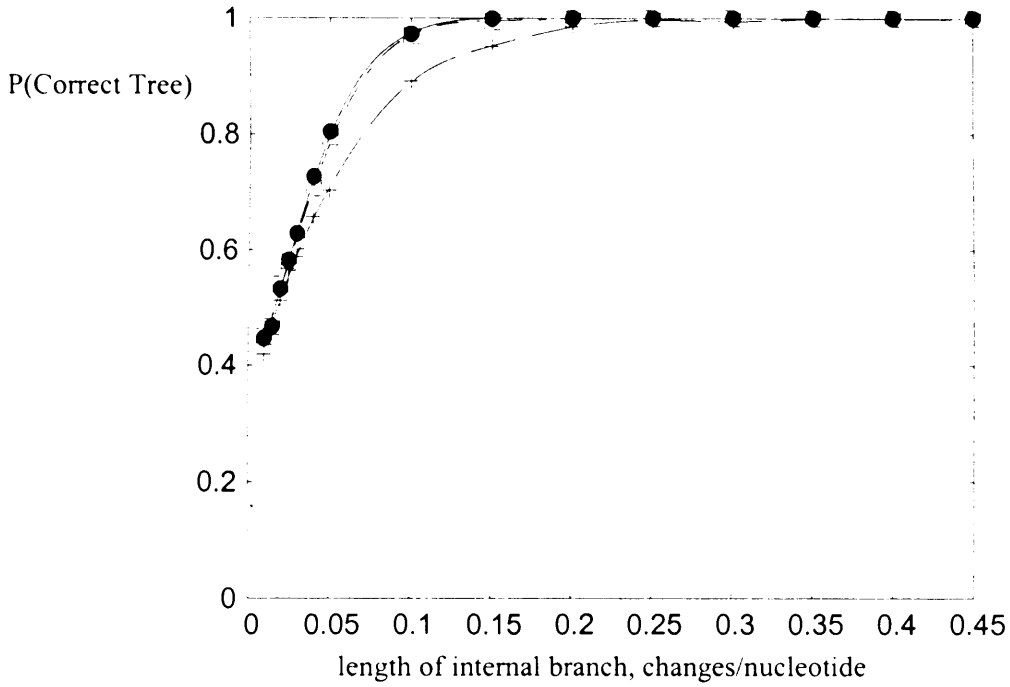


Figure 4.6. Showing the probability of reconstructing the correct tree plotted against the length of internal branch under a 5-class random branch simulation. The simulation conditions were heterotachous with 5 classes of branch lengths though the model used in reconstruction had only 2-classes of branch lengths. The internal branch lengths were the same for each class with external branch lengths drawn from a uniform distribution: U(0,5]. Phylogeny reconstruction was done using parsimony (solid line), heterotachous likelihood (dashed line) and traditional non-heterotachous likelihood (dotted line). The Jukes-Cantor model was used for both simulation and likelihood-based reconstruction.

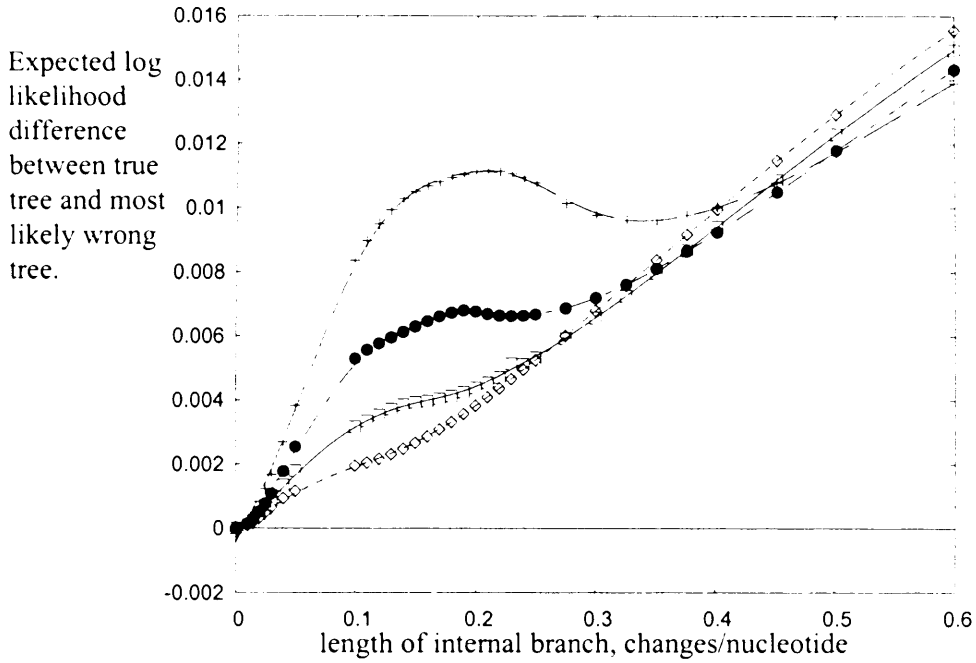


Figure 4.7. The expected log likelihood difference between the true tree and the most likely wrong tree when the true model and the model used for reconstructing the phylogeny are both heterotachous. The heterotachous conditions are similar to those in Figure 4.1, however the long and short external branch lengths are respectively: 0.35 and 0.45 (dotted lines/open diamonds), 0.60 and 0.20 (solid lines/ open squares), 0.70 and 0.10 (small dashes/filled dots), 0.75 and 0.05 (large dashes/ horizontal crosses). The Jukes-Cantor model was used for calculating both the probability of observing the data under the true tree topology and true heterotachous model and the probability inferred under the alternative topologies and non-heterotachous model.

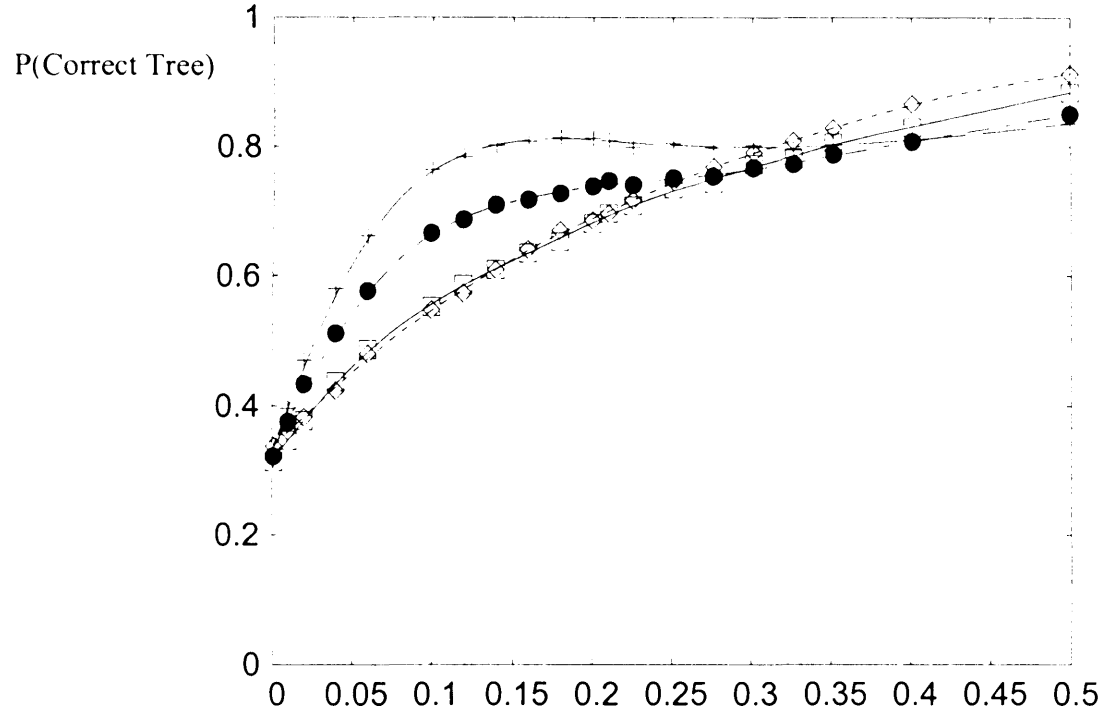


Figure 4.8. Showing the probability of reconstructing the correct tree versus length of internal branch with a 2-class heterotachy model used for simulation and reconstruction. The simulation conditions were heterotachous and similar to those of figure 4.1 but with different external branch lengths. These were: 0.35 and 0.45 (dotted lines/open diamonds), 0.60 and 0.20 (solid lines/ open squares), 0.70 and 0.10 (small dashes/filled dots), 0.75 and 0.05 (large dashes/ horizontal crosses). Jukes-Cantor was used for both simulation and likelihood-based reconstruction.

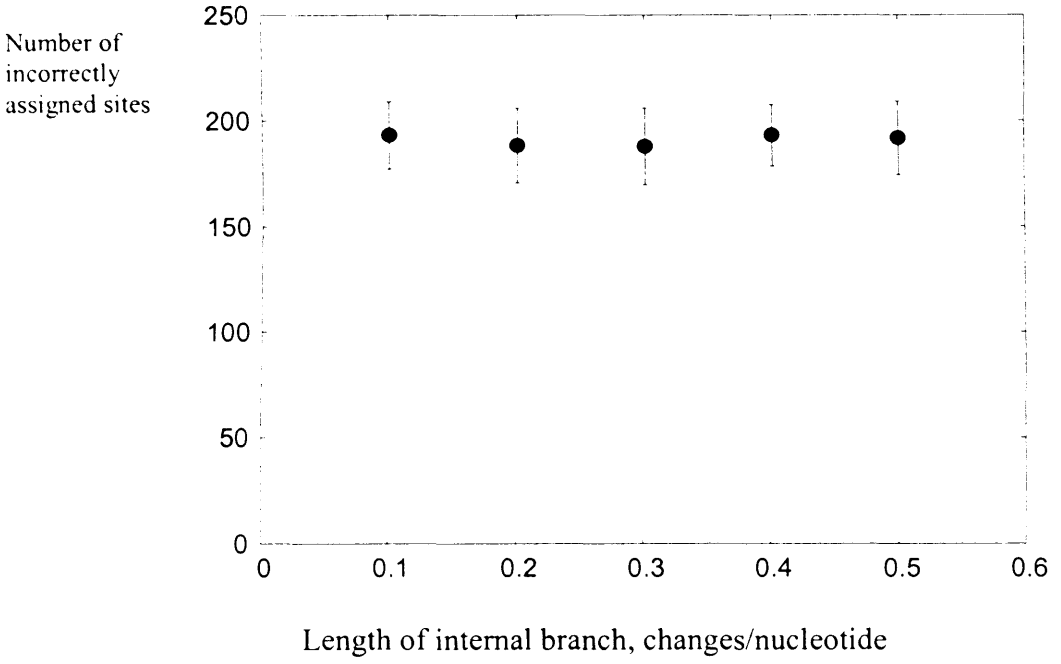


Figure 4.9. The mean numbers of sites incorrectly assigned to a heterotachy class. Sequence length is 500 nucleotides. Mean number shown by dots, twice the standard error by the bars

CHAPTER 5. COUNTING CHANGES PROBABILISTICALLY

5.1 Introduction

Direct comparisons between minimum change methods and statistical methods are problematic because the different types of method make inferences about evolution in different ways. The probabilistic models used in statistical methods define evolutionary processes in terms of instantaneous rates of change between states. The goal of the likelihood analysis is then to accurately estimate the parameters that determine those rates of change (WHELAN *et al.* 2001). Parsimony and other counting methods treat evolution as an accumulation of events that change the state of a character. The goal of the analysis is then to infer a minimal count of those events (FITCH 1971) that can account for the observed states. Patterns are detected based on the inferred counts of those changes. Attempts to reconcile the two approaches involve turning the counts inferred under a minimum-change criterion into rates for comparison with those inferred from a likelihood analysis. The rate calculated in this way has been compared to the rate inferred by a likelihood method in simulation studies, see for example (WONG *et al.* 2004).

The approach we take in this chapter is different in that to compare minimum change methods to likelihood based methods we do not turn the inferred count of changes into a rate. Instead we ask the question: if the rates of change defined under a probabilistic model are true, what is the probability of a count inferred using a minimum change method being correct? Thus instead of turning the counts of a minimum-change method into a rate, we turn the rates defined under a probabilistic model into a probability of number of changes made. In other words we turn the rate into a count. This method is applied to the

minimum change criterion models of (SUZUKI and GOJOBORI 1999), comparing them to probabilities predicted from the likelihood model (GOLDMAN and YANG 1994). Additionally we address the phylogeny reconstruction problem by comparing the counts of changes inferred by the parsimony method with the probability of that number of changes occurring in the 4-species case.

It can be seen that as well as providing a means of comparison between the two different frameworks, the method detailed here can provide a "multiple hit correction". Often analyses proceed by reconstructing ancestral states and counting changes that have been made between nodes, eg (AKASHI and VITINS 2005; TAKANO 1998). When one state is reconstructed at one end of a branch and a different one on the other, this is counted as one change. It is realised that this is inaccurate and the method described here allows for the assignment of a probability to the number of changes that could have occurred along the branch.

5.2 Methods

We consider a general likelihood model with k states. We denote the set of possible states as **X** and the i^{th} member of that set as x_i. We assume that we already have an instantaneous rate matrix, **Q**, with entries q_{ij} describing the instantaneous rate of transition between the states in **X**. We consider only a homogenous rate matrix that does not depend on time. There are $k(k-1)$ possible directed changes between those states. We denote an individual directed change as $(x_i \, T \, x_j)$ and the set of all possible changes as **C**. We are interested in a subset of those changes. Writing that subset as **S** we have $\mathbf{S} \in \mathbf{C}$. To give a concrete example, if we were interested in counting the number of non-synonymous changes in a codon sequence then the term x_i would represent a codon and **S** would comprise those single nucleotide changes that change the amino-acid encoded by the codon, e.g. (CCT T CTT).

5.2.1 The two-sequence case

The initial problem we are addressing is this: given the instantaneous rate matrix **Q** and a starting state x_i , what is the probability of the system ending up in state x_j after time t having made m changes of interest (i.e. changes that are members of **S**)?

We proceed by extending our definition of a state to include the number of changes made so far, and then constructing an instantaneous rate matrix that describes the rate of transition between the extended states. We define our extended state as the tuple $\{x_i, m\}$ where x_i is the state and m represents the number of changes of interest made since time t=0. It can be seen that the rate of transition from the extended state $\{x_i, m\}$ to the extended

state $\{x_j, n\}$, if we could write it out, could be represented by an infinite matrix, with rows and columns representing the states: $\{x_0, 0\}\ldots\{x_k, 0\},\{x_0, 1\}\ldots\{x_k, 1\},\{x_0, 2\}\ldots\{x_k, 2\}, \ldots$.

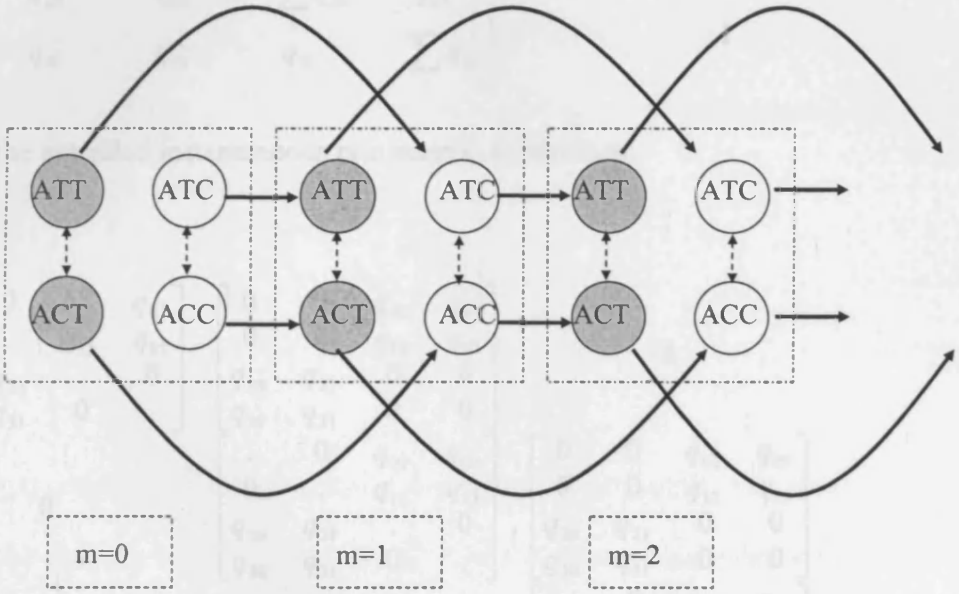
We denote this matrix **R**. This matrix can be divided into k x k submatrices. The submatrix occupying the mth row and nth column of the matrix **R** describe all the rates of transition between the set of states $\{x_0, m\}$ ... $\{x_k, m\}$ and the set of states $\{x_0, n\}$ ... $\{x_k, n\}$. We denote these submatrices $\mathbf{r_{m,n}}$. In our notation $[\mathbf{r_{m,n}}]_{ij}$ represents the $(i,j)^{th}$ entry of the submatrix $r_{m,n}$, ie the transition $\{x_i, m\}\mathrm{T}\{x_j, n\}$.

We define the instantaneous rate of more than one change occurring to be 0. Thus $\mathbf{r_{m,n}} = \mathbf{0}$ for $n \notin \{m, m+1\}$. If we view **R** as a block matrix whose entries are given by $\mathbf{r_{m,n}}$, all its entries are **0** except the matrices running along the diagonal and the superdiagonal of the block matrix. The former type of matrix represents the transitions $\{x_i, m\}$ T $\{x_j, m\}$, which we shall label $\mathbf{r_d}$, and the latter represents the transitions $\{x_i, m\}$ T $\{x_j, m+1\}$, which we shall label $\mathbf{r_{sd}}$.

It can be seen that $[\mathbf{r_d}]_{ij} = q_{ij}$ if the transition $(x_i\ \mathrm{T}\ x_j)$ is not a member of **S**, or $i=j$, and 0 otherwise. Equivalently, $[\mathbf{r_{sd}}]_{ij} = q_{ij}$ if the transition $(x_i\ \mathrm{T}\ x_j)$ is a member of **S** and 0 otherwise. One way to visualise this is to imagine the system, having made m changes, bouncing around making changes that are not of interest. The rates of these changes are determined by the submatrix $\mathbf{r_d}$. Eventually the system makes a change that is of interest and m increments to $m+1$. The rate of this second type of change is determined by $\mathbf{r_{sd}}$.

To give an explicit example, consider a reduced genetic code with 4 codons. In this example the possible state space **X** = {ATT, ATC, ACT, ACC}. Two of these codons, ATT and ATC, code for the amino acid isoleucine and the other two codons, ACT and ACC code for the amino acid threonine. If we are interested in calculating the probability of m

non-synonymous changes, the changes of interest are (ATTTACT) , (ATCTACC), (ACCTATC) and (ACTTATT). The possible transitions between the extended states can be represented by the diagram:



Isoleucine is shaded grey, threonine white. The number of non-synonymous changes made so far is given by m. Each box represents the set of extended states $\{x_i, m\}$ where x_i represents any possible codon. The system will start in the m=0 box, having made no changes of interest. It can bounce around in the m=0 box, possible changes represented by the dotted arrows. Eventually it will make a non-synonymous change, represented by the bold arrows and escape into the m=1 box. Here it will bounce around until it escapes into the m=2 box and so on. The rate of change within a box is determined by the submatrix $\mathbf{r}_d$ and the rate of change between boxes is represented by the submatrix $\mathbf{r}_{sd}$.

Numbering our states {ATT, ATC, ACT, ACC} as {0, 1, 2, 3} we can write the instantaneous rate matrix **Q** as:

$$\mathbf{Q} = \begin{bmatrix} -\sum_i q_{0i} & q_{01} & q_{02} & q_{03} \\ q_{10} & -\sum_i q_{1i} & q_{12} & q_{13} \\ q_{20} & q_{21} & -\sum_i q_{2i} & q_{23} \\ q_{30} & q_{31} & q_{32} & -\sum_i q_{3i} \end{bmatrix} \quad \text{.....(5.1)}$$

Thus the extended instantaneous rate matrix, **R** becomes:

$$\mathbf{R} = \begin{bmatrix} \begin{bmatrix} . & 0 & q_{02} & q_{03} \\ 0 & . & q_{12} & q_{13} \\ q_{20} & q_{21} & . & 0 \\ q_{30} & q_{31} & 0 & . \end{bmatrix} & \begin{bmatrix} 0 & 0 & q_{02} & q_{03} \\ 0 & 0 & q_{12} & q_{13} \\ q_{20} & q_{21} & 0 & 0 \\ q_{30} & q_{31} & 0 & 0 \end{bmatrix} & \mathbf{0} & \cdots \\ \mathbf{0} & \begin{bmatrix} . & 0 & q_{02} & q_{03} \\ 0 & . & q_{12} & q_{13} \\ q_{20} & q_{21} & . & 0 \\ q_{30} & q_{31} & 0 & . \end{bmatrix} & \begin{bmatrix} 0 & 0 & q_{02} & q_{03} \\ 0 & 0 & q_{12} & q_{13} \\ q_{20} & q_{21} & 0 & 0 \\ q_{30} & q_{31} & 0 & 0 \end{bmatrix} & \cdots \\ \mathbf{0} & \mathbf{0} & \begin{bmatrix} . & 0 & q_{02} & q_{03} \\ 0 & . & q_{12} & q_{13} \\ q_{20} & q_{21} & . & 0 \\ q_{30} & q_{31} & 0 & . \end{bmatrix} & \cdots \\ \mathbf{0} & \mathbf{0} & \mathbf{0} & \cdots \\ \cdots & \cdots & \cdots & \cdots \end{bmatrix}$$

.....(5.2)

Here the "." on the diagonal represents the sum of the terms on the rest of the row.

R is a denumerably infinite matrix describing the evolution of a Markov process in continuous time. It is stable, since the diagonal values are all equal to values from the matrix **Q** and hence are finite, and conservative since $-r_{ii} = \sum_{i \neq j} r_{ij}$ for all i. Again this property results from the derivation of **R** from **Q**.

Because **R** is stable, conservative and hence uniformisable (KIJIMA 1997), once we have the instantaneous rate matrix, we can solve the problem we are addressing by using the Kolmogorov forward equation:

$$P(t) = P(0)e^{\mathbf{R}t} \quad \text{.....(5.3)}$$

Here $P(t)$ is the vector of probabilities describing the probability that the system is in the extended state $\{x_i, m\}$ after time t, for all x_i and m. It can be seen that we must exponentiate **R**t. We denote the matrix $e^{\mathbf{R}t}$ as **E**, which we regard as a block matrix of k x k submatrices, in a similar manner to the way we regard **R**. The submatrix in **E** occupying an equivalent position to $[\mathbf{r}_{m,n}]_{ij}$ is written: $[\mathbf{e}_{m,n}]_{ij}$.

In the system we have described, the probability of making a transition $\{x_i, m\}$ T $\{x_j, m\}$ equals the probability of making the transition $\{x_i, m+1\}$ T $\{x_j, m+1\}$ and the probability of making the transition $\{x_i, m\}$ T $\{x_j, m+n\}$ in time t equals the probability of making the transition $\{x_i, m+1\}$ T $\{x_j, m+n+1\}$. Thus $\mathbf{e}_{m,n} = \mathbf{e}_{m+1,n+1}$. Therefore to calculate E, we are only need calculate the top row of submatrices of $e^{\mathbf{R}t}$, i.e $\mathbf{e}_{0,n}$.

$$\exp(Rt) = I + Rt + \frac{R^2t^2}{2!} + \frac{R^3t^3}{3!} + \text{.....(5.4)}$$

If we limit ourselves to diagonalizable models for **Q**, we can set $\mathbf{r_d} = \mathbf{UDU^{-1}}$, and factorize **R** so that the superdiagonal becomes $\mathbf{U^{-1}}\ \mathbf{r_{sd}}\ \mathbf{U}$.

Writing this explicitly we have:

$$\mathbf{R} = \begin{bmatrix} \mathbf{U} & \mathbf{0} & \mathbf{0} & \cdots \\ \mathbf{0} & \mathbf{U} & \mathbf{0} & \cdots \\ \mathbf{0} & \mathbf{0} & \mathbf{U} & \cdots \\ \cdots & \cdots & \cdots & \cdots \end{bmatrix} \cdot \begin{bmatrix} \mathbf{D} & \mathbf{U^{-1}r_{sd}U} & \mathbf{0} & \cdots \\ \mathbf{0} & \mathbf{D} & \mathbf{U^{-1}r_{sd}U} & \cdots \\ \mathbf{0} & \mathbf{0} & \mathbf{D} & \cdots \\ \cdots & \cdots & \cdots & \cdots \end{bmatrix} \cdot \begin{bmatrix} \mathbf{U^{-1}} & \mathbf{0} & \mathbf{0} & \cdots \\ \mathbf{0} & \mathbf{U^{-1}} & \mathbf{0} & \cdots \\ \mathbf{0} & \mathbf{0} & \mathbf{U^{-1}} & \cdots \\ \cdots & \cdots & \cdots & \cdots \end{bmatrix} \quad \text{.....(5.5)}$$

Noting the upper-diagonal form of the matrix, and writing $\mathbf{M} = \mathbf{U^{-1}}\ \mathbf{r_{sd}}\ \mathbf{U}$, we define the recursive formula:

$$\begin{aligned} \mathbf{R}^{1}_{0,0} &= \mathbf{U.D^{1}.U^{-1}} \\ \mathbf{R}^{1}_{0,1} &= \mathbf{U.M.U^{-1}} \\ \mathbf{R}^{1}_{0,j} &= \mathbf{0};\ \mathbf{j} \notin \{\mathbf{0,1}\} \\ \mathbf{R}^{n}_{0,j} &= \mathbf{U.(D.R^{n-1}_{0,j} + R^{n-1}_{0,j-1}.M).U^{-1}} \end{aligned} \quad \text{.....(5.6)}$$

Here $\mathbf{R}^{w}_{n,m}$ refers to the k x k matrix on the n^{th} row and m^{th} column of the block matrix: **R** raised to the power w. Using the recursion algorithm we can calculate the entries in the sum of equation 4.2 and hence we can calculate **E**. In practice a scaling and squaring step is included to improve numerical accuracy. Also, though in principle an infinite number of changes are possible, in practice the probability of a large number of changes over a biologically reasonable branch length becomes very small, and we limit the number of k x k submatrices in **E** and **R** accordingly.

Once **E** has been calculated, $P(t)$ can be calculated using equation 4.1.

5.2.2 The many sequence case

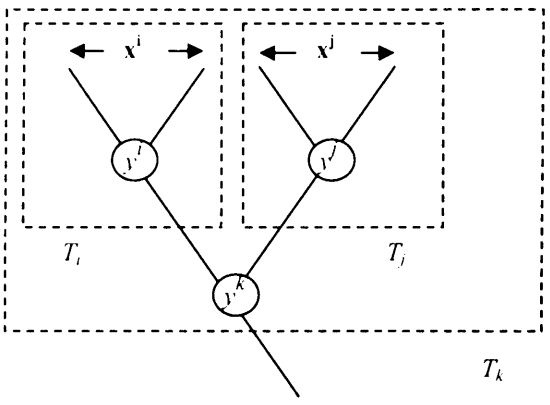
So far we have shown how to turn an instantaneous rate matrix, **Q**, and a starting state, x_i, into a probability distribution describing the probability of the system making m changes of interest in time t, and ending up in state x_j. We write this probability as:

$$P(x_j, n \mid x_i ; t) \quad \text{.....(5.7)}$$

In the many sequence case we have many sequences connected by a tree. States at the tips are known, but states at internal nodes are not and have to be summed over. In this section we describe an algorithm for calculating the probability of a system making m changes of interest over a tree and ending up with the data observed at the tips. To do this we will describe a recursion step that is an extension of Felsenstein's pruning algorithm (FELSENSTEIN 1981) and then describe the initiation and termination steps.

Recursion

If we consider a case where a subtree T_k is a subtree of a larger tree T and two sub trees T_i and T_j are subtrees of T_k connected by the node y^k.



$\mathbf{x}^{\mathbf{i}}$ is here the data observed at all the tips above the node y^i, and equivalently $\mathbf{x}^{\mathbf{j}}$ represents all the data observed at the tips above the node y^j. We assume for the purposes of the description of the recursive step the following condition. We assume that we know the probability of the system making n_i changes of interest over subtree T_i and ending up with data $\mathbf{x}^{\mathbf{i}}$, given that the state at y^i is known. We write the u^{th} state at the i^{th} node as y^i_u. We assume that this calculation has been performed for each y^i_u and for each number of changes n_i. We write this as:

$$P(\mathbf{x}^{\mathbf{i}}, n_i \mid y^i_u) \quad \text{.....(5.8)}$$

Equivalently we assume for T_j that we have already calculated, for each state at y^j, and number of changes n_j:

$$P(\mathbf{x}^{\mathbf{j}}, n_j \mid y^j_u) \quad \text{.....(5.9)}$$

To calculate the equivalent probailities for T_k we proceed in 2 stages. Firstly we calculate $P(\mathbf{x}^{\mathbf{i}}, n_i \mid y^k_v)$ and $P(\mathbf{x}^{\mathbf{j}}, n_j \mid y^k_v)$ using the formulas:

$$P(\mathbf{x}^{\mathbf{i}}, n_i \mid y^k_v) = \sum_{r=0}^{n_i} \sum_u P(\mathbf{x}^{\mathbf{i}}, n_i - r \mid y^i_u) P(y^i_u, r \mid y^k_v) \quad \text{.....(5.10)}$$

$$P(\mathbf{x}^{\mathbf{j}}, n_j \mid y^k_v) = \sum_{r=0}^{n_j} \sum_u P(\mathbf{x}^{\mathbf{j}}, n_j - r \mid y^j_u) P(y^j_u, r \mid y^k_v) \quad \text{.....(5.11)}$$

The inner summation is over the possible states at the nodes y^i and y^j respectively. It can be seen that the first term on the right-hand side of equation (5.10) is known from (5.8). The second term on the right-hand side can be calculated from equation (5.7) describing the two sequence case. Once this has been calculated the second stage is to calculate:

$$P(\mathbf{x}^i, \mathbf{x}^j, n_k \mid y_v^k) = \sum_r P(\mathbf{x}^i, n_k - r \mid y_v^k) P(\mathbf{x}^j, r \mid y_v^k) \quad \text{.....(5.12)}$$

The terms on the right hand side have already been calculated in steps (5.10) and (5.11).

Initiation

The initiation step is relatively straightforward in that the data at the tips are observed. Thus writing x_c^k for the site at external node k observed in state c we can write:

$$P(x_c^k, n_k \mid y_v^k) = \begin{cases} 1; n_k = 0 \wedge v = c \\ 0; otherwise \end{cases} \quad \text{.....(5.13)}$$

At the tip, no changes above the tip node can have occurred, hence n_k has to equal 0.

It can be seen that we can now recursively pass down the tree. Initially the tips are set, as described in the initiation step. Then their parents are visited and set as described in the recursion step. Once the parents are set, their parents are visited and set and so on down the tree. At the root we perform the termination step.

Termination

Assuming a trifurcating root the termination step is given by:

$$P(\mathbf{x}^{\mathbf{i}},\mathbf{x}^{\mathbf{j}},\mathbf{x}^{\mathbf{k}},n_r)=\sum_{u}\sum_{\upsilon=0}^{n_k}P(\mathbf{x}^{\mathbf{i}},n_k-\nu\mid y_u^0)\sum_{\nu=0}^{\upsilon}P(\mathbf{x}^{\mathbf{j}},\upsilon-\nu\mid y_u^0)P(\mathbf{x}^{\mathbf{k}},\nu\mid y_u^0).P(u)$$

.....(5.14)

Where the root node in state u is written y_u^0 and the equilibrium probability of the state u is written $P(u)$.

This algorithm describes how to calculate the probability of a system making n changes of interest over a tree and ending up with the data **x** at the tips. In general we will be interested in calculating the probability of the system making n changes over the tree, *given that* it has ended up with data **x** at the tips. To do this we calculate:

$$P(n\mid\mathbf{x})=\frac{P(\mathbf{x},n)}{\sum_{n}P(\mathbf{x},n)}$$

.....(5.15)

5.2.3 Numerical computation

The algorithm as described raises some practical difficulties. Firstly the probabilities being multiplied and summed become very small leading to numerical inaccuracies. To obviate these, we note that:

$$P(\mathbf{x}^{\mathbf{i}}, n_i \mid y_v^k) = \frac{1}{a_i} \sum_{r=0}^{n_i} \sum_u a_i P(\mathbf{x}^{\mathbf{i}}, n_i - r \mid y_u^i) P(y_u^i, r \mid y_v^k) \quad \text{.....(5.16)}$$

And:

$$P(\mathbf{x}^{\mathbf{j}}, n_j \mid y_v^k) = \frac{1}{a_j} \sum_{r=0}^{n_j} \sum_u a_j P(\mathbf{x}^{\mathbf{j}}, n_j - r \mid y_u^j) P(y_u^j, r \mid y_v^k) \quad \text{.....(5.17)}$$

Thus we can scale the parent node accordingly:

$$P(\mathbf{x}^{\mathbf{i}}, \mathbf{x}^{\mathbf{j}}, n_k \mid y_v^k) = \frac{a_k}{a_i . a_j} \sum_r a_i P(\mathbf{x}^{\mathbf{i}}, n_k - r \mid y_v^k) a_j P(\mathbf{x}^{\mathbf{j}}, r \mid y_v^k) \quad \text{.....(5.18)}$$

Hence as we recurse down the tree we keep track of both the term $a_k P(\mathbf{x}^{\mathbf{k}}, n_k \mid y^k{}_v)$ and the scaling term $log\ (a_k)$. The scaling term is chosen so that the values of n_k and v that maximise the term $P(\mathbf{x}^{\mathbf{k}}, n_k \mid y^k{}_v)$ are scaled so that the largest value of the expression $a_k P(\mathbf{x}^{\mathbf{k}}, n_k \mid y^k{}_v)$ equals 1.0. From the standpoint of the numerical computation, this ensures that the largest terms, ie the ones that contribute most to the sum described in equation (5.10) are calculated most accurately. This, however, comes at the expense of calculating the smaller terms inaccurately. An alternative strategy would be to scale $a_k P(\mathbf{x}^{\mathbf{k}}, n_k \mid y^k{}_v)$ such that the mean value was 1.0. However the increase in accuracy gained in the smallest terms is more than offset by the loss of accuracy in the larger terms.

The second practical difficulty is similar to the one described in the two sequence case. In principle an infinite number of changes are possible. In practice we limit the number of changes along a branch. This in turn allows for a method to speed-up the algorithm. As before we consider the topology T with subtree T_k which has subtrees T_i and T_j joined by node y^k. We allow only n_{ki}* changes on the branch from y^k to y^i and n_{kj}* changes on the branch from y^k to y^j. Assuming we know the maximum number of changes permitted in the tree above the nodes y^i and y^j, which we denote n_i* and n_j* respectively, we can calculate the maximum number of changes at node y^k as:

$$n_k^* = n_i^* + n_j^* + n_{ki}^* + n_{kj}^*. \quad \text{.....(5.19)}$$

We initiate this algorithm for finding the maximum number of allowed changes by noting that the maximum number of changes at a tip is 0.

To speed the calculation up we note that when calculating $P(\mathbf{x}^i, n_i \mid y^k_v)$, as in equation (5.10), the sum only has to be performed for each value of n_i up to $n_i^* + n_{ki}^*$, and when calculating $P(\mathbf{x}^k, n_k \mid y^k_v)$ as in equation (5.12), the sum only has to be performed for values of n_k up to n_k^*. In practice n_k^* can still get very big for large trees. However the larger values of n_k^* have very low probabilities so we set a maximum number of changes of interest over the whole tree denoted, n_{max}.

5.2.4 Model validation

To validate the model in the two sequence case over a time period t, we simulated a poisson jump process. A state was chosen at random, labelled x_i. The waiting time for that state was determined by a random number, a_0, drawn from an exponential distribution with mean waiting time determined by the entry q_{ii} in the rate matrix, **Q**. If a_0 was greater than t, we terminated, else we picked another state. We picked the state x_j ($j \neq i$) with probability:

$$P(x_j) = \frac{q_{ij}}{\sum_{k \neq i} q_{ik}} \quad \text{.....(5.20)}$$

If the change (x_iTx_j) was a change of interest we incremented the count of changes of interest by one. Once the change had been made we again allowed the system to remain in state x_j, this time the waiting time a_1, taken from an exponential distribution with mean $-q_{jj}$. If $a_0+a_1 < t$ we repeated the process until $\sum_i a_i > t$.

Thus over a time period t we were able to count the number of changes of interest made. This was repeated and the observed probabilities of the system making n changes of interest compared to the expected probability of the system making n changes of interest, calculated as described. The comparison was done using a χ^2 test.

To verify the many sequence case, the probability of a site making n changes of interest was calculated as:

$$P(n) = \sum_{\mathbf{x}} P(\mathbf{x}, n) \quad \text{.....(5.21)}$$

To calculate (5.21) we note that:

$$\sum_{\mathbf{x}^i}\sum_{\mathbf{x}^j} P(\mathbf{x}^i, \mathbf{x}^j, n_k \mid y_v^k) = \sum_{r}\sum_{\mathbf{x}^i} P(\mathbf{x}^i, n_k - r \mid y_v^k)\sum_{\mathbf{x}^j} P(\mathbf{x}^j, r \mid y_v^k) \quad \text{.....(5.22)}$$

Then we repeat the algorithm described in section (5.2.1) with a different initiation step At a tip node y^k:

$$P(x_c^k, n_k \mid y_v^k) = \begin{cases} 1; n_k = 0 \\ 0; otherwise \end{cases} \quad \text{.....(5.23)}$$

Once this had been calculated we repeated the poisson jump simulation previously described, starting at the root and proceeding down the tree assigning states to internal nodes when the waiting time at a state exceeded the branch length. The observed number of changes of interest were compared to the proportion expected.

As a further test, data was simulated using the jump process described. At the tips the final state of the process was recorded as was the number of changes of interest made over the tree. We write the data at the tips as $\mathbf{d_{obs}}$ and the number of changes of interest recorded as n_{obs}. We then used (5.15) to calculate:

$$P(n > n_{obs} \mid \mathbf{d_{obs}}) = \sum_{n_i \geq n_{obs}}^{n_{max}} P(n_i \mid \mathbf{d_{obs}}) \quad \text{.....(5.24)}$$

We then repeated the process resimulating new values for n_{obs} and $\mathbf{d_{obs}}$, counting how often the probability in (5.24) fell into the region 0-0.05, 0.05-0.10 ... 0.95-1.00. These counts were tested against the counts expected from a uniform distribution using a χ^2 test.

5.2.5 Application to comparisons with parsimony

To calculate the probability of parsimony methods correctly reconstructing the number of changes over a tree we used the Jukes-Cantor and HKY85 model on a four sequence tree. The expected probability of the number of counts inferred under parsimony being correct is given as:

$$P(n_p) = \sum_{\mathbf{x}} P(\mathbf{x}).P(n_p(\mathbf{x}) \mid \mathbf{x}) \quad \text{.....(5.25)}$$

Where $n_p(\mathbf{x})$ is the count of changes inferred under the parsimony criterion when the data is $\mathbf{x}$. This probability was calculated for the 4-species tree shapes that were described in Chapter 4, table 4.1. In the HKY85 simulation the base frequencies were set at 0.4, 0.3, 0.2, 0.1 for the bases T, C, A, G respectively. The transition/transversion ratio was set at 5.0.

To test the accuracy of the counts of non-synonymous changes and synonymous changes inferred at a single site using the method of (SUZUKI and GOJOBORI 1999), two studies were performed. In the first, a 30-species tree was selected from the PANDIT database(WHELAN *et al.* 2003), shown in Figure 5.4. This was scaled, keeping the ratio of the branch lengths the same but changing the total tree length. Data was simulated on the tree using the codon model of (GOLDMAN and YANG 1994). The codon frequencies were set at equality and the transition / transversion rate ratio was set at 5.0. The value of ω, the nonsynonymous / synonymous rate ratio was set at 1.0. 500 codon sites were simulated at each total tree length and the numbers of synonymous and non-synonymous changes were counted using the method of Suzuki and Gojobori. The probability of this being the correct

number of changes was then calculated using the algorithm described using the true model to construct the **R** matrix of equation (5.5). Additionally the probability of the correct number being less than or equal to the count generated by the cladistic method was addressed. The question addressed by this study is: given that we know the actual model, what is the probability of the method of Suzuki and Gojobori correctly counting the number of changes.

The second test was to investigate whether this was a problem that real tree shapes incurred. Tree shapes of 10, 20, 30, 40, 50 and 60 sequences were taken from the PANDIT database. For each tree shape an analysis similar to the one described was performed. Data was simulated on the tree using a codon model with a single nonsynonymous/synonymous rate ratio, set at 1.0. The codon frequencies were again set at equality and the transition / transversion rate ratio was set at 5.0. 100 codon sites were simulated for tree and the number of synonymous and non-synonymous changes was counted using the method of Suzuki and Gojobori. The probability of this count being less than or equal to the actual number of changes, given the true model, was then calculated as before.

5.2.6 Application to the "multiple hit correction"

To demonstrate the efficacy of the method applied to the multiple hit problem, the cumulative probability of n changes between 2 sequences, given that a site is in state T at one end and G at the other was calculated. The model used was HKY85 (HASEGAWA *et al.* 1985) with a transition transversion ratio of 5.0 and base frequencies of 0.4, 0.3, 0.2, 0.1 for the bases T, C, A, G respectively.

5.3 Results and discussion

The results of Figure 5.1 show the probability that the count of changes derived under the parsimony criterion is correct, given that the true model is that of Jukes-Cantor and the tree shapes are as shown. It can be seen that at large divergences, the probability that the count of changes derived under a parsimony criterion is correct becomes very low. It seems therefore that methods which use the count of changes derived under the parsimony criterion should be used with caution at large divergences. It can also be seen that there is a relation between the probability of reconstructing the correct tree in a phylogenetic analysis and the probability of the number of changes under the parsimony criterion being correct. The parsimony-derived count has a lower probability of being correct under the Felsenstein tree of figure 1 than it does under the Farris tree. The Felsenstein tree is harder to reconstruct by parsimony in a phylogenetic analysis.

In Figure 5.1, we also see the probability that the count of changes derived under the parsimony criterion is correct given that the true model is HKY85. Again, perhaps unsurprisingly, the probability of the count of changes derived under the parsimony criterion being correct becomes low at high divergences. The probability of the parsimony-derived count being correct is marginally smaller when the true model is HKY85 than it is when the true model is Jukes-Cantor. This is expected because the parsimony method being used, Fitch parsimony, weights all changes as being equally likely. This seems intuitively closer to the Jukes-Cantor model. Indeed it has been argued that parsimony is the same as the Jukes-Cantor method with short branch lengths (JUKES 1969). However the effect is small, it seems that the probability of the parsimony derived account being correct depends

far more on the distance between the sequences than it does on the correctness of the model.

Figure 5.2 shows the probability of the inferred counts of synonymous and nonsynonymous changes at a site being correct when the Suzuki and Gojobori method is used and the true model is one of neutral evolution, with the tree shown in figure 5.3. It can be seen that the probability of the count being correct again becomes very low, even when the total tree length is relatively small. A total tree length of 15 changes per codon is equivalent to one change per six nucleotides per sequence. Additionally the bias in the count of synonymous changes is greater than that of the non-synonymous changes. Thus if the count of synonymous changes is relatively low, then the estimates of the non-synonymous/synonymous ratio will be biased upwards. This difference occurs because there are different numbers of possible non-synonymous and synonymous changes. Figure 5.4 shows a similar result when real 30 sequence trees are taken from the PANDIT database. Again the count of nonsynonymous changes has a much higher probability of being correct than the count of synonymous changes for trees whose branch lengths are derived from real biological sequences.

Figure 5.5 shows the result of the calculation of the probability of n changes between 2 nucleotides at the distances shown, given the states at the end of the branch are T and G. Again it can be seen that at divergences even as low as 0.3 changes / nucleotide, a simple minimum change counting method will only count correctly about 80% of the time.

5.3 Conclusion

It conclusion it can be shown that this method accurately turns rates defined under a statistical method into a distribution that describes the probability of the system making a given number of changes of interest. The description of the method is sufficiently broad that it can not only be applied to counting the number of changes a system makes, but also to counting the number of a subset of all possible changes. In this chapter we applied the method both to calculating the probability that a nucleotide makes n changes of any type over a phylogeny and to calculating the probability of a given number of non-synonymous and synonymous changes happening over a phylogeny.

Counts derived from minimum change methods have a low probability of being correct at high divergences. This is a well understood phenomenon and this method may provide a "multiple hit correction" to counting methods (AKASHI and VITINS 2005). It is perhaps surprising how low the probability of a correct count gets even at small divergences. However it is also noticeable that the count of non-synonymous and synonymous changes are not equally likely to be correct, which may lead to problems for methods based on counting an inferred number of non-synonymous and synonymous mutations.

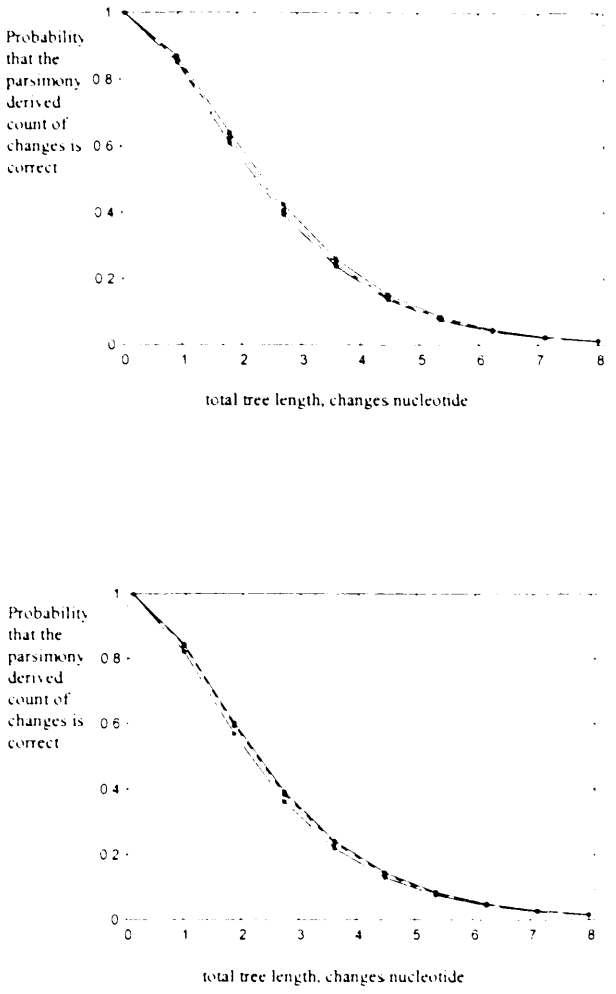
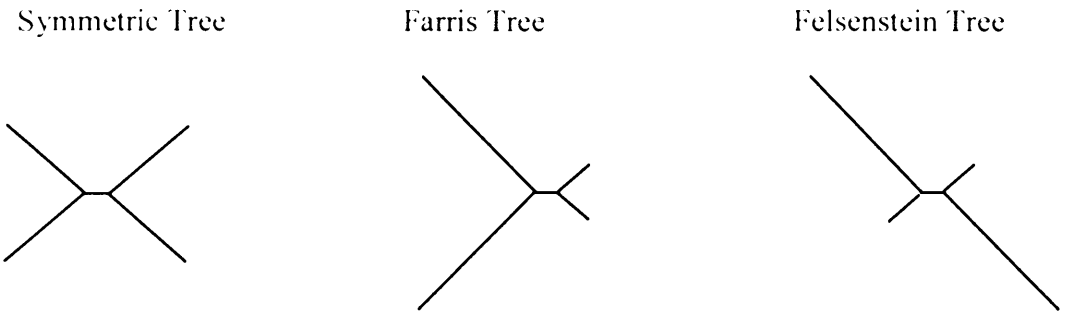


Figure 5.1. Probability that the count of changes derived under the parsimony criterion is correct at different total tree lengths, when the true model is Jukes Cantor (top) and HKY85 (bottom). The tree shapes are shown below. Results for the symmetric tree are shown by solid lines, the Farris tree by dotted lies and the Felsenstein tree by dashed lines.



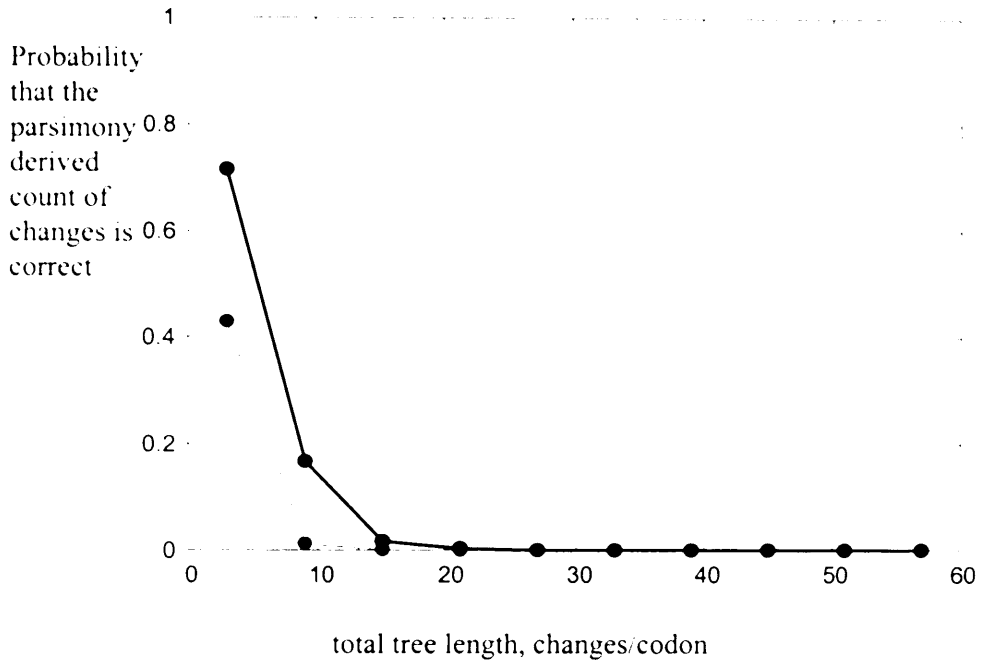


Figure 5.2. Probability that the count of nonsynonymous (solid line) and synonymous (dotted line) is correct when made using the Suzuki and Gojobori method using the topology and relative branch lengths shown in figure 5.4.

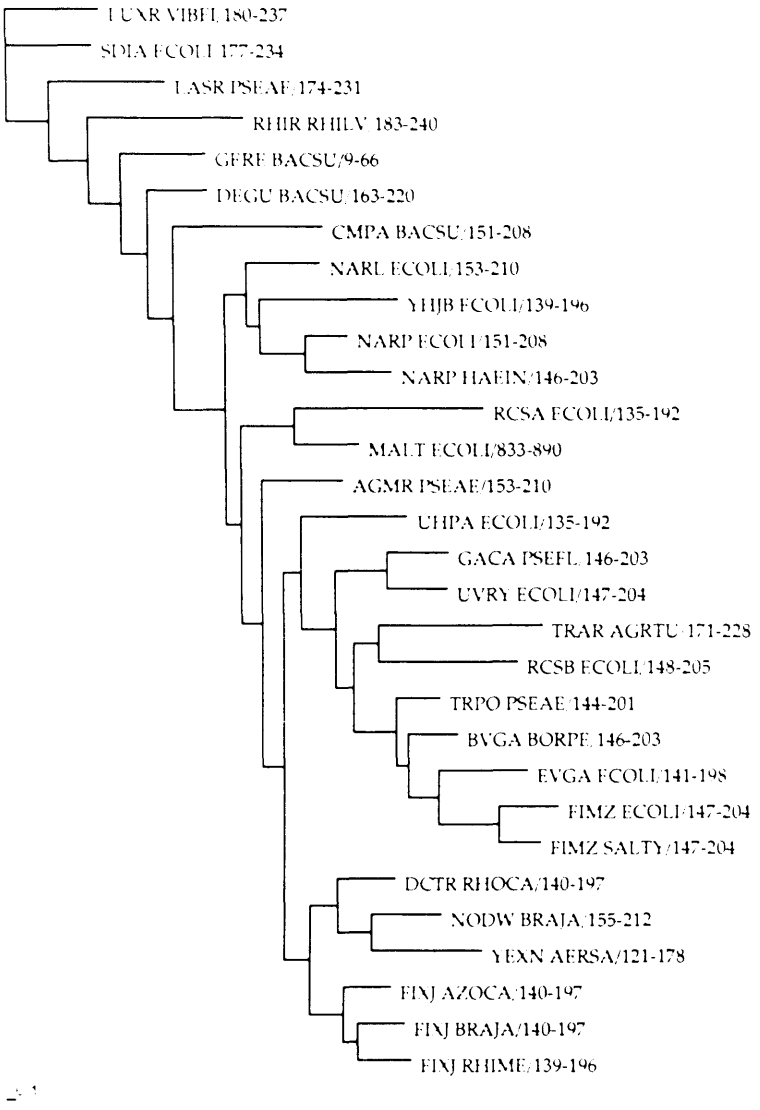


Figure 5.3 The tree topology and relative branch lengths used in the simulation

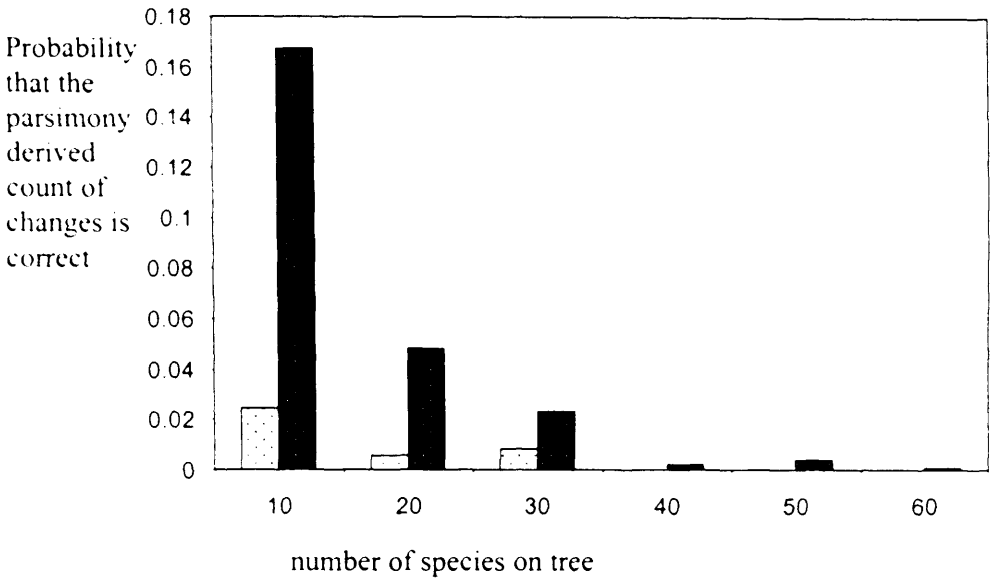


Figure 5.4. Probability that the count of nonsynonymous (solid bars) and synonymous (white bars) is correct when made using the Suzuki and Gojobori method using the trees of 10, 20, 30, 40 50 and 60 sequences present in the PANDIT database.

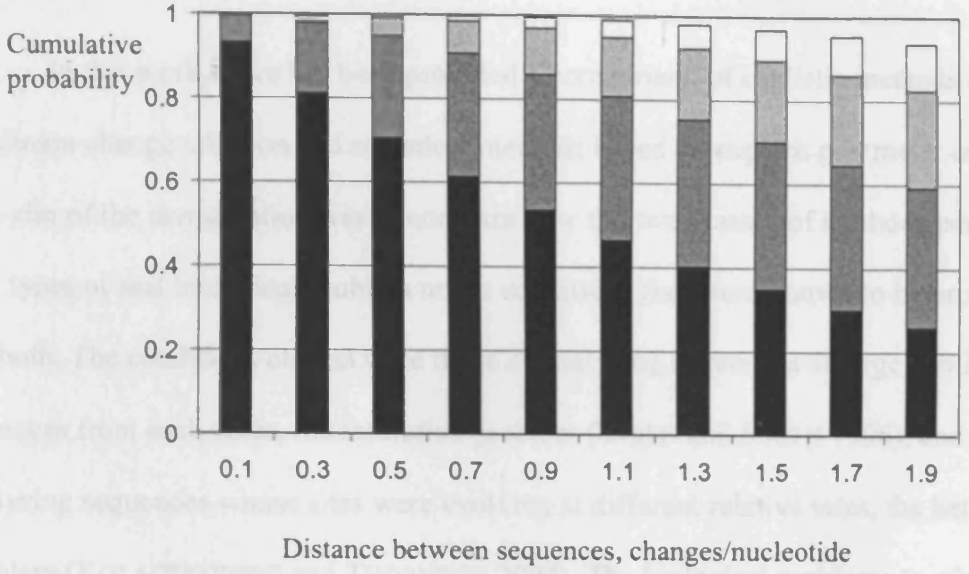


Figure 5.5. Cumulative probability of *n* changes between 2 sequences, given that a site is in state T at one end and G at the other. The model used is HKY85 with a transition transversion ration of 5.0 and base frequencies of 0.4, 0.3, 0.2, 0.1 for the bases T.C.A.G respectively. The probability of 1 change is shown by the black bars, 2 changes by the dark grey bars, 3 changes by the light grey bars and 4 changes by the white bars.

PERSPECTIVES

In this work, there has been provided a comparison of cladistic methods based on a minimum change criterion and statistical methods based on explicit parameter estimation. The aim of the investigation was to compare how the two classes of methods performed on two types of real biological problem under conditions that were known to be problematic for both. The conditions chosen were those of analysing sequences at large evolutionary distances from each other, the saturation problem (SMITH and SMITH 1996), and that of analysing sequences whose sites were evolving at different relative rates, the heterotachy problem (KOLACZKOWSKI and THORNTON 2004). The biological problems to which these methods were applied were reconstructing the evolutionary relationship between sequences and differentiating between adaptive and neutral evolution.

It seems that under the conditions of the study, statistical methods are a better choice for detecting adaptive evolution at large evolutionary distances. Additionally they are better for detecting a bias in the transition/transversion ratio. Even though the statistical methods do show some bias in the mean estimate of parameter values when the sequence length is small, this effect is negligible compared to the bias in the mean estimate of the minimum change methods at high divergences. Additionally though the variances in the parameter estimate are mostly comparable, where the minimum change methods have a lower variance they also have a strongly biased mean estimate. Thus one can be very sure of the wrong estimate. Finally when hypothesis testing with the parametric statistical methods, the expected type I error rate is close to the observed type I error rate even at large evolutionary distances. This is not true of the minimum change methods, which show a collapse in performance. The very different performance of the minimum change methods

and the parametric statistical methods leads to the proposition that stauration occurs by two different mechanisms. Minimum change methods cease to work because the minimum change criterion is violated, whilst parametric statistical methods cease to work because the probability that a site is observed in a state on one sequence becomes independent of its state on the other sequence(s).

Whilst for large sequences there is a good fit between the mathematically expected decline in performance of the parametric methods and the observed decline, this is not true at smaller sequence lengths. This, together with the observed bias in the mean, suggests an avenue of further work. The theory quoted in chapter 3 relies on a first order expansion of the score vector. However by examining higher order expansions of the expected likelihood surface, one can both correct the bias in the mean estimate and more accurately describe the distribution of the maximum-likelihood estimates. For example one could use the p^*-formula (BARNDORFF-NIELSEN 1994):

$$p^*(\hat{\theta};\theta_0 \mid a) = c\left|\hat{j}\right|^{\frac{1}{2}} \exp(\ell(\theta_0;\hat{\theta},a) - \ell(\hat{\theta};\hat{\theta},a)) \quad \text{.....(P.1)}$$

Here $\hat{j}$ is the observed information matrix, θ is the parameter of interest, a is an ancillary statistic and c is chosen so that the total integral of equation (C.1) is 1.0. The problems with the use of this formula include actually performing the integral and the use of an appropriate ancillary statistic. In practice, it would seem most fruitful to use an approximately ancillary statistic, for example the directed-likelihood. This kind of approach might be most useful when investigating selection on very short peptides (YANG and SCHEPARTZ 2005).

Moving from the detection of adaptive evolution to the problem of tree-reconstruction, it does not seem possible to draw a similar conclusion about the relative performance of minimum-change methods and parametric statistical methods. In the cases studied, both minimum change methods and parametric statistical methods showed an approximately similar decline in performance. The exception to this was the difference in performance between the methods when investigating the "Felsenstein tree" and the "Farris tree". The different effect of saturation on the relative performance of the two classes of method on tree reconstruction and detection of adaptive evolution gives an indication that, as previously suggested (YANG *et al.* 1995a), phylogeny estimation is not a problem of parameter estimation but of model definition. This may explain why minimum change methods, interpreted as reconstructing the most probable evolutionary history, seem effective in most cases at phylogeny reconstruction but are less effective when the phylogeny is known and a parameter needs to be estimated. Further work may elucidate this difference more clearly.

It has been shown that heterotachous evolution is less of a problem for parsimony than for non-heterotachous likelihood models in a specific evolutionary scenario with a specific combination of branch lengths. However the superiority of parsimony in that evolutionary scenario disappears if the simulation model is made more realistic. If branch lengths are drawn from a random distribution, both the parsimony method and the non-heterotachous likelihood method performed approximately equivalently well. Again this can be explained if the tree topology is not a statistical parameter in the typical sense. Hence parsimony by choosing the most likely evolutionary history, including ancestral node states, might be expected to perform equivalently to likelihood.

The heterotachous likelihood model performed unexpectedly badly, actually doing worse in certain situations than the incorrect non-heterotachous likelihood model. Additionally it performed poorly at correctly assigning sites to different rate classes. This is of interest because similar models have been the subject of recent study (GADAGKAR and KUMAR 2005; SPENCER *et al.* 2005). It is possible that the poor performance was due to the large number of parameters present in the likelihood analysis. An alternative approach might be to construct an empirical Bayesian model in which the prior probability of each branch in each class being a given length is drawn from a gamma distribution whose parameters are estimated using the method of maximum likelihood. The probability of the data could be calculated by summing over the possible branch lengths using a Markov Chain Monte Carlo approach, see for example (RONQUIST and HUELSENBECK 2003) . Alternatively it may be that in real sequences groups of sites contiguous on the sequence are likely to be in the same heterotachy class. This has the advantage of biological realism since in a protein coding region of DNA, sites that define a single functional motif are likely to be clustered together and be under similar evolutionary pressures. It might be of use in that case to define an auto-correlative model similar to the rate models (FELSENSTEIN and CHURCHILL 1996; YANG 1995).

Finally a method was presented for comparing the counts of changes inferred under a minimum change analysis to the rates inferred under a parametric statistical analysis. This was done by turning the rates into a probability distribution describing the probability of a number of changes of interest, given the data. This then allowed us to calculate the probability of the count under the minimum change analysis being correct. Perhaps unsurprisingly at large divergences the probability that the counts derived under a minimum

change criterion were correct became very low. This suggests that methods that use counts of changes as a starting point eg (AKASHI and VITINS 2005; TAKANO 1998) should be used with caution at high divergences. This method may be able to make a "multiple-hit correction" so that these methods will work at larger distances. Also it was shown that the parsimony derived count had a higher probability of being correct for the Farris tree, known to be easy for parsimony to reconstruct correctly, than for the Felsenstein tree, known to be difficult for parsimony to reconstruct correctly.

The rates-to-count approach seems to provide an adequate basis for the comparison of minimum change methods and likelihood methods. However it might be possible to use the distribution of changes for inference. One possible question is: given a non-synonymous/synonymous rate ratio of 1.0, can I reasonably explain my data without inferring an unexpectedly high number of non-synonymous changes? Unfortunately it seems that the answer to this question is always yes. It seems that it is the relative number of non-synonymous and synonymous changes that is important. For this approach to be useful it might be possible to infer simultaneously the probability of x non-synonynmous changes and y synonymous changes having occurred, given the data.

Thus in conclusion it seems that though minimum-change methods may provide an adequate means for phylogenetic estimation they should be used with caution when inferring patterns of evolution, especially at high divergences. It seems that under these conditions likelihood methods perform better.

REFERENCES

AKASHI, H., KO, W, PIAO, S, ANOOP JOHN, PIYUSH GOEL, CHIAO-FENG LIN,, and A. P. VITINS, 2005 Molecular evolution in the Drosophila melanogaster species subgroup: Frequent parameter fluctuations on the time-scale of molecular divergence. In preparation.

AMARI, S. I., 1987 Estimation of structural parameter in the presence of infinitely many nuisance paramters, pp. 73-83 in *Differential Geometry in Statistical Inference*, edited by S. I. AMARI, BARNDORFF-NIELSEM, O.E., KASS, R.E., LAURITZEN, S.L., RAO, C.R. Instiute of Mathematical Statistics.

ANDREWS, C. W., 1904 On the evolution of the Proboscidea. Philosophical Transactions of the Royal Society of London. Series B, Containing Papers of a Biological Character **196:** 99-198.

ARNDT, P. F., D. A. PETROV and T. HWA, 2003 Distinct changes of genomic biases in nucleotide substitution at the time of Mammalian radiation. Mol Biol Evol **20:** 1887-1896.

AXELSSON, E., M. T. WEBSTER, N. G. SMITH, D. W. BURT and H. ELLEGREN, 2005 Comparison of the chicken and turkey genomes reveals a higher rate of nucleotide divergence on microchromosomes than macrochromosomes. Genome Res **15:** 120-125.

BARNDORFF-NIELSEN, O. E., AND COX, D.R., 1994 *Inference and Asymptotics*. Chapman and Hall.

BARRY, D. H., J.A., 1987 Statistical Analysis of Hominoid Molecular Evolution. Statistical Science **2:** 191-207.

BENTON, M. J., 1985 Classification and Phylogeny of the Diapsid Reptiles. Zoological Journal of the Linnean Society **84:** 97-164.

BOFFELLI, D., M. A. NOBREGA and E. M. RUBIN, 2004 Comparative genomics at the vertebrate extremes. Nat Rev Genet **5:** 456-465.

BROCCHIERI, L., 2001 Phylogenetic inferences from molecular sequences: review and critique. Theor Popul Biol **59:** 27-40.

CAL, S., and B. A. CONNOLLY, 1997 DNA distortion and base flipping by the EcoRV DNA methyltransferase. A study using interference at dA and T bases and modified deoxynucleosides. J Biol Chem **272:** 490-496.

CAMIN, J. H., and R. R. SOKAL, 1965 A Method for Deducing Branching Sequences in Phylogeny. Evolution **19:** 311-326.

CANBACK, B., I. TAMAS and S. G. ANDERSSON, 2004 A phylogenomic study of endosymbiotic bacteria. Mol Biol Evol **21:** 1110-1122.

CAVALLI-SFORZA, L. L. A. E., A.W.F., 1967 Phylogenetic analysis: models and estimation procedures. Evolution **21:** 550-570.

CHANG, J. T., 1996 Inconsistency of evolutionary tree topology reconstruction methods when substitution rates vary across characters. Math Biosci **134:** 189-215.

CLARKE, B., 1970 Darwinian evolution of proteins. Science **168:** 1009-1011.

CRAMER, H., 1946 *Mathematical Methods of Statistics*. Princeton University Press, Princeton.

CRICK, F., 1970 Central dogma of molecular biology. Nature **227:** 561-563.

DARWIN, C., 1859 *The Origin of Species*

by Means of Natural Selection.

DAYRAT, B., 2003 The roots of phylogeny: how did Haeckel build his trees? Syst Biol **52:** 515-527.

DE BAERE, E., D. BEYSEN, C. OLEY, B. LORENZ, J. COCQUET *et al.*, 2003 FOXL2 and BPES: mutational hotspots, phenotypic variability, and revision of the genotype-phenotype correlation. Am J Hum Genet **72:** 478-487.

DOMAZET-LOSO, T., and D. TAUTZ, 2003 An evolutionary analysis of orphan genes in Drosophila. Genome Res **13:** 2213-2219.

DURBIN, R., EDDY, S., KROGH, A., MITCHISON, G., 1998 *Biological sequence analysis.* Cambridge University Press.

EDWARDS, A. W. F., 1992 *Likelihood.* Johns Hopkins University Press.

FARRIS, J. S., 1970 Methods for Computing Wagner Trees. Systematic Zoology **19:** 83-&.

FARRIS, J. S., 1973 Probability Model for Inferring Evolutionary Trees. Systematic Zoology **22:** 250-256.

FELSENSTEIN, J., 1978 Cases in Which Parsimony or Compatibility Methods Will Be Positively Misleading. Systematic Zoology **27:** 401-410.

FELSENSTEIN, J., 1981 Evolutionary trees from DNA sequences: a maximum likelihood approach. J Mol Evol **17:** 368-376.

FELSENSTEIN, J., 1983 Statistical-Inference of Phylogenies. Journal of the Royal Statistical Society Series a-Statistics in Society **146:** 246-272.

FELSENSTEIN, J., 1985 Phylogenies and the Comparative Method. American Naturalist **125:** 1-15.

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

FITCH, W. M., 1971 Toward Defining Course of Evolution - Minimum Change for a Specific Tree Topology. Systematic Zoology **20:** 406-&.

FLEISSNER, R., D. METZLER and A. VON HAESELER, 2005 Simultaneous statistical multiple alignment and phylogeny reconstruction. Syst Biol **54:** 548-561.

FRIEDMAN, R., J. W. DRAKE and A. L. HUGHES, 2004 Genome-wide patterns of nucleotide substitution reveal stringent functional constraints on the protein sequences of thermophiles. Genetics **167:** 1507-1512.

GADAGKAR, S. R., and S. KUMAR, 2005 Maximum Likelihood Outperforms Maximum Parsimony Even When Evolutionary Rates Are Heterotachous. Mol Biol Evol **22:** 2139-2141.

GALTIER, N., 2001 Maximum-Likelihood Phylogenetic Analysis Under a Covarion-like Model. Mol Biol Evol **18:** 866-873.

GILAD, Y., O. MAN and G. GLUSMAN, 2005 A comparison of the human and chimpanzee olfactory receptor gene repertoires. Genome Res **15:** 224-230.

GILL, P. E., MURRAY, W., AND WRIGHT, M.H., 1981 *Practical Optimization*. Academic Press, San Diego.

GOLDMAN, N., 1990 Maximum-Likelihood Inference of Phylogenetic Trees, with Special Reference to a Poisson-Process Model of DNA Substitution and to Parsimony Analyses. Systematic Zoology **39:** 345-361.

GOLDMAN, N., 1998 Phylogenetic information and experimental design in molecular systematics. Proc Biol Sci **265:** 1779-1786.

GOLDMAN, N., and Z. YANG, 1994 A codon-based model of nucleotide substitution for protein-coding DNA sequences. Mol Biol Evol **11:** 725-736.

GROSOVSKY, A. J., J. G. DE BOER, P. J. DE JONG, E. A. DROBETSKY and B. W. GLICKMAN, 1988 Base substitutions, frameshifts, and small deletions constitute ionizing radiation-induced point mutations in mammalian cells. Proc Natl Acad Sci U S A **85:** 185-188.

GU, Z., Y. ZHANG, P. SHI, Y. P. ZHANG, D. ZHU *et al.*, 2004 Comparison of avian myostatin genes. Anim Genet **35:** 470-472.

GUINDON, S., and O. GASCUEL, 2003 A simple, fast, and accurate algorithm to estimate large phylogenies by maximum likelihood. Systematic Biology **52:** 696-704.

HAECKEL, E., 1866 *Generelle Morphologie der Organismen.* Reimer, Berlin.

HASEGAWA, M., H. KISHINO and T. YANO, 1985 Dating of the human-ape splitting by a molecular clock of mitochondrial DNA. J Mol Evol **22:** 160-174.

HENNIG, W., 1966 *Phylogenetic Systematics, [Translated by Davis D.D. & Zangerl R. from Hennig W. 1950. Grundzüge einer Theorie der Phylogenetischen Systematik. Deutscher Zentralverlag, Berlin.].* Uni. Illinois Press, Urbana.

HILLIER, L. W., W. MILLER, E. BIRNEY, W. WARREN, R. C. HARDISON *et al.*, 2004 Sequence and comparative analysis of the chicken genome provide unique perspectives on vertebrate evolution. Nature **432:** 695-716.

HOLMES, S., 2003 Statistics for phylogenetic trees. Theoretical Population Biology **63:** 17-32.

HUELSENBECK, J. P., 2002 Testing a covariotide model of DNA substitution. Mol Biol Evol **19:** 698-707.

HUELSENBECK, J. P., and K. A. CRANDALL, 1997 Phylogeny estimation and hypothesis testing using maximum likelihood. Annual Review of Ecology and Systematics **28:** 437-466.

HUGHES, A. L., and R. FRIEDMAN, 2005 Variation in the pattern of synonymous and nonsynonymous difference between two fungal genomes. Mol Biol Evol **22:** 1320-1324.

INA, Y., 1998 Estimation of the transition/transversion ratio. J Mol Evol **46:** 521-533.

INAGAKI, Y., E. SUSKO, N. M. FAST and A. J. ROGER, 2004 Covarion shifts cause a long-branch attraction artifact that unites microsporidia and archaebacteria in EF-1alpha phylogenies. Mol Biol Evol **21:** 1340-1349.

JANVIER, P., 1996 The dawn of the vertebrates: Characters versus common ascent in the rise of current vertebrate phylogenies. Palaeontology **39:** 259-287.

JEFFREYS, T., 1961 *Theory of Probability*. Oxford University Press.

JUKES, T. A. C., C, 1969 Evolution of protein molecules in *Mammalian protein metabolism*, edited by H. MUNRO. Academic Press, New York.

KIJIMA, M., 1997, pp. 183-199 in *Markov Processes for Stochastic Modelling*. Chapman and Hall.

KIMURA, M., 1968 Evolutionary rate at the molecular level. Nature **217:** 624-626.

KIMURA, M., 1980 A simple method for estimating evolutionary rates of base substitutions through comparative studies of nucleotide sequences. J Mol Evol **16:** 111-120.

KIMURA, M., 1983 *The Neutral Theory of Molecular Evolution*. Cambridge University Press.

KING, J. L., and T. H. JUKES, 1969 Non-Darwinian evolution. Science **164:** 788-798.

KLUGE, A. G., 2001 Philosophical conjectures and their refutation. Systematic Biology **50:** 322-330.

KLUGE, A. G., and J. S. FARRIS, 1969 Quantitative Phyletics and Evolution of Anurans. Systematic Zoology **18:** 1-&.

KOLACZKOWSKI, B., and J. W. THORNTON, 2004 Performance of maximum parsimony and likelihood phylogenetics when evolution is heterogeneous. Nature **431:** 980-984.

KREUZER, K. N., 2005 Interplay between DNA replication and recombination in prokaryotes. Annu Rev Microbiol **59:** 43-67.

LEMMON, A. R., and M. C. MILINKOVITCH, 2002 The metapopulation genetic algorithm: An efficient solution for the problem of large phylogeny estimation. Proc Natl Acad Sci U S A **99:** 10516-10521.

LEQUESNE, W. J., 1974 Uniquely Evolved Character Concept and Its Cladistic Application. Systematic Zoology **23:** 513-517.

LEWONTIN, R. C., and J. KRAKAUER, 1973 Distribution of gene frequency as a test of the theory of the selective neutrality of polymorphisms. Genetics **74:** 175-195.

LI, W. H., C. I. WU and C. C. LUO, 1985 A new method for estimating synonymous and nonsynonymous rates of nucleotide substitution considering the relative likelihood of nucleotide and codon changes. Mol Biol Evol **2:** 150-174.

LI, Y., Y. P. QIAN, X. J. YU, Y. Q. WANG, D. G. DONG *et al.*, 2004 Recent origin of a hominoid-specific splice form of neuropsin, a gene involved in learning and memory. Mol Biol Evol **21:** 2111-2115.

LOCKHART, P. J., M. A. STEEL, A. C. BARBROOK, D. H. HUSON, M. A. CHARLESTON *et al.*, 1998 A covariotide model explains apparent phylogenetic structure of oxygenic photosynthetic lineages. Mol Biol Evol **15:** 1183-1188.

LOPEZ, P., D. CASANE and H. PHILIPPE, 2002 Heterotachy, an important process of protein evolution. Mol Biol Evol **19:** 1-7.

LYNN, D. J., G. A. SINGER and D. A. HICKEY, 2002 Synonymous codon usage is subject to selection in thermophilic bacteria. Nucleic Acids Res **30:** 4272-4277.

MACINTYRE, G., C. V. ATWOOD and C. G. CUPPLES, 2001 Lowering S-adenosylmethionine levels in Escherichia coli modulates C-to-T transition mutations. J Bacteriol **183:** 921-927.

MADDISON, D. R. A. K.-S. S. E., 2004 The Tree of Life Web Project. Internet address: http: tolweb.org, pp.

MAERE, S., S. DE BODT, J. RAES, T. CASNEUF, M. VAN MONTAGU *et al.*, 2005 Modeling gene and genome duplications in eukaryotes. Proc Natl Acad Sci U S A **102:** 5454-5459.

MAKI, H., 2002 Origins of spontaneous mutations: specificity and directionality of base-substitution, frameshift, and sequence-substitution mutageneses. Annu Rev Genet **36:** 279-303.

MESSIER, W., and C. B. STEWART, 1997 Episodic adaptive evolution of primate lysozymes. Nature **385:** 151-154.

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

MIYATA, T., and T. YASUNAGA, 1980 Molecular evolution of mRNA: a method for estimating evolutionary rates of synonymous and amino acid substitutions from homologous nucleotide sequences and its application. J Mol Evol **16:** 23-36.

MURATA-KAMIYA, N., H. KAMIYA, N. IWAMOTO and H. KASAI, 1995 Formation of a mutagen, glyoxal, from DNA treated with oxygen free radicals. Carcinogenesis **16:** 2251-2253.

MUSE, S. V., and B. S. GAUT, 1994 A likelihood approach for comparing synonymous and nonsynonymous nucleotide substitution rates, with application to the chloroplast genome. Mol Biol Evol **11:** 715-724.

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

NIELSEN, R., and Z. YANG, 1998 Likelihood models for detecting positively selected amino acid sites and applications to the HIV-1 envelope gene. Genetics **148:** 929-936.

OHTA, T., 2002 Near-neutrality in evolution of genes and gene regulation. Proc Natl Acad Sci U S A **99:** 16134-16137.

OWEN, A., 2001 *Empirical Likelihood*. Taylor and Francis.

PAPOULIS, A., 1984 *Probability, Random Variables, and Stochastic Processes*. New York: McGraw-Hill.

PHILIPPE, H., D. CASANE, S. GRIBALDO, P. LOPEZ and J. MEUNIER, 2003 Heterotachy and functional shift in protein evolution. IUBMB Life **55:** 257-265.

PHILIPPE, H., and J. LAURENT, 1998 How good are deep phylogenetic trees? Curr Opin Genet Dev **8:** 616-623.

PUPKO, T., and N. GALTIER, 2002 A covarion-based method for detecting molecular adaptation: application to the evolution of primate mitochondrial genomes. Proc Biol Sci **269:** 1313-1316.

REDELINGS, B. D., and M. A. SUCHARD, 2005 Joint Bayesian estimation of alignment and phylogeny. Syst Biol **54:** 401-418.

RICHMOND, R. C., 1970 Non-Darwinian evolution: a critique. Nature **225:** 1025-1028.

RIVERO, F., T. MURAMOTO, A. K. MEYER, H. URUSHIHARA, T. Q. UYEDA *et al.*, 2005 A comparative sequence analysis reveals a common GBD/FH3-FH1-FH2-DAD architecture in formins from Dictyostelium, fungi and metazoa. BMC Genomics **6:** 28.

RONQUIST, F., and J. P. HUELSENBECK, 2003 MrBayes 3: Bayesian phylogenetic inference under mixed models. Bioinformatics **19:** 1572-1574.

RUIZ-TRILLO, I., J. PAPS, M. LOUKOTA, C. RIBERA, U. JONDELIUS *et al.*, 2002 A phylogenetic analysis of myosin heavy chain type II sequences corroborates that Acoela and Nemertodermatida are basal bilaterians. Proc Natl Acad Sci U S A **99:** 11246-11251.

SIDDALL, M. E., 1998 Success of parsimony in the four-taxon case: Long-branch repulsion by likelihood in the Farris Zone. Cladistics-the International Journal of the Willi Hennig Society **14:** 209-220.

SMITH, J. M., and N. H. SMITH, 1996 Synonymous nucleotide divergence: what is "saturation"? Genetics **142:** 1033-1036.

SPENCER, M., E. SUSKO and A. J. ROGER, 2005 Likelihood, parsimony, and heterogeneous evolution. Mol Biol Evol **22:** 1161-1164.

STEEL, M., and D. PENNY, 2000 Parsimony, likelihood, and the role of models in molecular phylogenetics. Molecular Biology and Evolution **17:** 839-850.

STRAZEWSKI, P., 1988 Mispair formation in DNA can involve rare tautomeric forms in the template. Nucleic Acids Res **16:** 9377-9398.

SULLIVAN, J., and D. L. SWOFFORD, 2001 Should we use model-based methods for phylogenetic inference when we know that assumptions about among-site rate variation and nucleotide substitution pattern are violated? Syst Biol **50:** 723-729.

SUZUKI, Y., and T. GOJOBORI, 1999 A method for detecting positive selection at single amino acid sites. Molecular Biology and Evolution **16:** 1315-1328.

TAKANO, T. S., 1998 Rate variation of DNA sequence evolution in the Drosophila lineages. Genetics **149:** 959-970.

TAKEZAKI, N., and M. NEI, 1994 Inconsistency of the Maximum Parsimony Method When the Rate of Nucleotide Substitution Is Constant. Journal of Molecular Evolution **39:** 210-218.

TAMURA, K., and M. NEI, 1993 Estimation of the number of nucleotide substitutions in the control region of mitochondrial DNA in humans and chimpanzees. Mol Biol Evol **10:** 512-526.

THOMPSON, E. A., 1975 *Human evolutionary trees*. Cambridge University Press.

TUFFLEY, C., and M. STEEL, 1997 Links between maximum likelihood and maximum parsimony under a simple model of site substitution. Bulletin of Mathematical Biology **59:** 581-607.

WALLIS, O. C., A. O. MAC-KWASHIE, G. MAKRI and M. WALLIS, 2005 Molecular evolution of prolactin in primates. J Mol Evol **60:** 606-614.

WHELAN, S., P. I. W. DE BAKKER and N. GOLDMAN, 2003 Pandit: a database of protein and associated nucleotide domains with inferred trees. Bioinformatics **19:** 1556-1563.

WHELAN, S., P. LIO and N. GOLDMAN, 2001 Molecular phylogenetics: state-of-the-art methods for looking into the past. Trends Genet **17:** 262-272.

WONG, W. S., Z. YANG, N. GOLDMAN and R. NIELSEN, 2004 Accuracy and power of statistical methods for detecting adaptive evolution in protein coding sequences and for identifying positively selected sites. Genetics **168:** 1041-1051.

YANG, L., and A. SCHEPARTZ, 2005 Relationship between folding and function in a sequence-specific miniature DNA-binding protein. Biochemistry **44:** 7469-7478.

YANG, Z., 1993 Maximum-likelihood estimation of phylogeny from DNA sequences when substitution rates differ over sites. Mol Biol Evol **10:** 1396-1401.

YANG, Z., 1994a Estimating the pattern of nucleotide substitution. J Mol Evol **39:** 105-111.

YANG, Z., 1994b Maximum likelihood phylogenetic estimation from DNA sequences with variable rates over sites: approximate methods. J Mol Evol **39:** 306-314.

YANG, Z., 1995 A space-time process model for the evolution of DNA sequences. Genetics **139:** 993-1005.

YANG, Z., 1996 Phylogenetic analysis using parsimony and likelihood methods. Journal of Molecular Evolution **42:** 294-307.

YANG, Z., 1997 PAML: a program package for phylogenetic analysis by maximum likelihood. Computer Applications in the Biosciences **13:** 555-556.

YANG, Z., 1998a Likelihood ratio tests for detecting positive selection and application to primate lysozyme evolution. Mol Biol Evol **15:** 568-573.

YANG, Z., 1998b On the best evolutionary rate for phylogenetic analysis. Syst Biol **47:** 125-133.

YANG, Z, 2000a Complexity of the simplest phylogenetic estimation problem. Proceedings of the Royal Society of London Series B-Biological Sciences **267:** 109-116.

YANG, Z., 2000b Maximum likelihood estimation on large phylogenies and analysis of adaptive evolution in human influenza virus A. Journal of Molecular Evolution **51:** 423-432.

YANG, Z., N. GOLDMAN and A. FRIDAY, 1995a Maximum-Likelihood Trees from DNA-Sequences - a Peculiar Statistical Estimation Problem. Systematic Biology **44:** 384-399.

YANG, Z., S. KUMAR and M. NEI, 1995b A new method of inference of ancestral nucleotide and amino acid sequences. Genetics **141:** 1641-1650.

YANG, Z., and R. NIELSEN, 2000 Estimating synonymous and nonsynonymous substitution rates under realistic evolutionary models. Mol Biol Evol **17:** 32-43.

YANG, Z., W. S. WONG and R. NIELSEN, 2005 Bayes empirical bayes inference of amino acid sites under positive selection. Mol Biol Evol **22:** 1107-1118.

ZHANG, J., 2004 Frequent false detection of positive selection by the likelihood method with branch-site models. Mol Biol Evol **21:** 1332-1339.

ZHANG, J., R. NIELSEN and Z. YANG, 2005 Evaluation of an Improved Branch-Site Likelihood Method for Detecting Positive Selection at the Molecular Level. Mol Biol Evol **22:** 2472-2479.

ZHANG, J., H. F. ROSENBERG and M. NEI, 1998 Positive Darwinian selection after gene duplication in primate ribonuclease genes. Proc Natl Acad Sci U S A **95:** 3708-3713.

ZHARKIKH, A., and W. H. LI, 1993 Inconsistency of the Maximum-Parsimony Method - the Case of 5 Taxa with a Molecular Clock. Systematic Biology **42:** 113-125.

ZHENG, D., Z. ZHANG, P. M. HARRISON, J. KARRO, N. CARRIERO *et al.*, 2005 Integrated pseudogene annotation for human chromosome 22: evidence for transcription. J Mol Biol **349:** 27-45.