The Parametric Bootstrap in Phylogenetic Analysis

A thesis submitted in partial fulfilment of the

requirements of the degree of

Masters of Science in Biological Sciences

The University of Auckland

New Zealand

February 2004

0.1 Abstract

The field of evolutionary inference from molecular sequences is currently a burgeoning field that requires strict standards as well as pioneering ideas to fully interpret the plethora of information being generated in molecular biology. Phylogenetic methods attempt to resolve evolutionary histories from the molecular data. The glut of available information poses new challenges to the field because in many instances the data do not all support the same hypotheses (topologies). This thesis is a simulation-based study of phylogenetic methods that assesses error rates in some of the methods of topology hypothesis testing, and provides an exploration of real data for developing new methods for model testing and model selection. Overall, the number of simulation runs presented in this thesis totals in the hundreds of millions, including simulation of replicate data sets, resampling of sequence data to generate pseudoreplicates and phylogenetic estimation from these replicate data sets. It involves the use of many currently available software packages and the analysis of these results with a number of Perl scripts. My results demonstrate that different tests, both essentially designed to test for a significant difference in topologies, yield answers that do not always agree. The reasons behind this and the implications for systematists wanting to use these tests are discussed. In addition, the type I error of a widely used test, the Swofford-Olsen-Waddell-Hillis test (SOWH test), was estimated and shown to be excessive under conditions of model violation. The strict and correct use of these tests still remains an issue with significantly more simulation work still to be done. Finally, the site-patterns of real pre-aligned data sets were analysed using a parametric bootstrapping approach that allows one to compare models and to estimate hypotheses about the patterns of evolution amongst the taxa.

0.2 Acknowledgements

Eight months ago, about half way through this thesis, when asked whom I might acknowledge, it never really dawned on me how much you really appreciate the little things that make a year of commitment to a single objective possible. So, at this time, I look back at the help I did receive from all avenues and am very grateful to you all. I have certainly depended on numerous people for support and general moral boosts to see it through and so I acknowledge everyone that contributed to the success of this year with overwhelming gratitude.

For supervising this thesis and for providing the direction, advice and editorial comments required to meet the standards set by the University of Auckland and the scientific community as a whole, I thank Professor Allen Rodrigo and Dr. Thomas Buckley.

For providing assistance in various matters relating to phylogenetics and computational biology in general, I thank Howard Ross, Greg Ewing, Matt Goode, and Stephane Guindon.

For gladly giving up time to proofread some of the material in this thesis, I thank Matthew Barrett and Bronwen Jongbloed.

For believing in me since day one, I thank my parents, Elsie and Liston Meintjes.

For breaking up the day on a consistent basis with visits to my office, I thank Richard Bunker and Eddie Walker.

For the Experience of the year, I thank Matt Barrett, Jacinta Alexander and Richard Bunker.

For keeping me sane by taking me for on/off campus lunches and beer, I thank Matt Barrett, Sean Tobin, Stu Preece, Hayden Smith, Alistair Law, Rob Thomas and Aashish Patel.

For a healthy bout of exercise on Sundays, I thank Julie Harper for organising "The Team".

For winning the Super 12 trophy, I thank the Auckland Blues.

For winning the Ranfurly Shield and the NPC, I thank the Auckland A Team.

For week in and week out supporting Auckland Rugby and providing an unparalleled camaraderie, I thank the Terrace Choir; Aaashish Patel, Chirag 'Chico' Chhita, Nic 'Da C' Francis, Alistair Law, James 'George' McKearney, Kamal Meria and Rob Thomas

For making it possible to study and earn a living simultaneously by teaching piano at Lewis Eady, I thank Matt Shanks, Teresa Cooper and Emi Steedman and John Eady.

0.3 Table of Contents

0.4 Terminology and Definitions

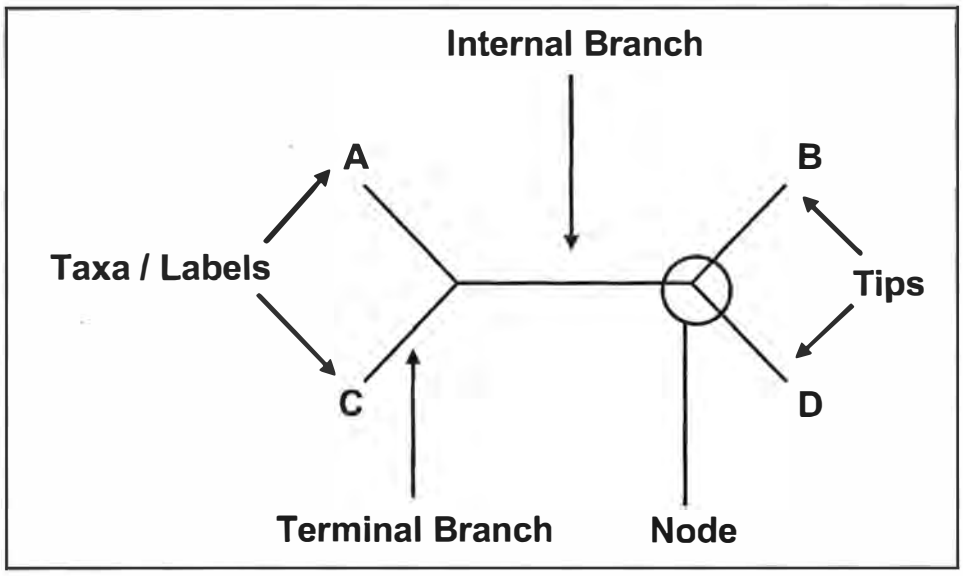

Figure 0.1 The terminology on an unrooted tree.

Consistency: The property of a statistical estimator to converge to the true value as more and more data are added.

Convergent evolution: See Homoplasy

δ: The SOWH test statistic for the difference in log likelihoods between the two evolutionary hypotheses, τ_0 and τ_1.

δ`: An estimate of the true value for δ, the difference in log likelihoods for two competing evolutionary hypotheses, calculated from parametric replicate data when the true tree and true model are known.

Δ_i: The difference in log likelihoods between two hypotheses, τ_0 and τ_1, for any parametric replicate data set. Collectively all Δ_i values comprise the Δ distribution.

Δ distribution: The null distribution of differences in log likelihoods between the null and alternative hypotheses for all the parametric replicates.

External branch lengths: The distance from a tip to the closest node

Farris tree: A tree exhibiting 'long-branch repulsion' between two related taxa on a four taxon tree. It is also known as the anti-Felsenstein zone tree (Waddell 1995) and the inverse-Felsenstein zone tree (Swofford et al., 2001).

Felsenstein tree: A tree exhibiting long-branch attraction between two unrelated taxa on a four taxon tree.

Homoplasy: The independent acquisition of the same nucleotide in unrelated taxa at a given site, such that it appears to be a homologous nucleotide.

Horizontal transfer: The transfer of genetic material between organisms without reproduction.

Internal branch lengths: The distance along a branch between two internal nodes.

Likelihood ratio test (LRT): The ratio of likelihoods as a statistical test for the goodness-of-fit between two models. The precise statistic is $2\ln(L(\theta_1)-L(\theta_2)$.

Maximum likelihood (ML): In phylogenetics maximum likelihood is an optimality criterion that selects the tree on which the observed sequence data has the highest probability of evolving.

Maximum parsimony: An optimality criterion that minimises the number of character state changes needed to explain the observed molecular sequence data.

Neighbour-Joining: A distance-based greedy algorithm for building trees from molecular sequences.

Non-synonymous substitution: A mutation at the nucleotide level that alters the amino acid for which that codon codes.

Outgroup: A taxon used to to root a tree that is hypothesised to be more diverged from all the other sequences than they are from one another.

Parallel evolution: See Homoplasy

Rate matrix: The square 4 x 4 matrix of relative instantaneous rates of nucleotide substitution.

Reassortment: The reorganisation of gene order within a genome.

Recombination: The physical exchange of DNA through crossing over.

Site-Pattern: A single column in a multiple alignment of homologous sequences that partitions the taxa in a particular way.

Split: A pattern that partitions the set of taxa into two or more non-empty subsets.

Synonymous substitution: A mutation at the nucleotide level that does not alter the amino acid for which that codon codes.

Topology: Defined as the shape of the tree or the evolutionary hypothesis for the taxa in the data set.

Transitions: The change of a nucleotide from a purine to a purine OR a pyrimidine to a pyrimidine (e.g. A←→ G or C←→T).

Transversions: The change of a nucleotide from a purine to a pyrimidine and vice versa (e.g. A→C or T, G→C or T, T→A or G, and C→A or G).

Tree: A visual representation of the sequential separation of taxa into distinct lineages which do not rejoin. It consists of the topology, τ, and the branch lengths l.

Type I error: The probability of rejecting the null hypothesis when it is true.

Type II error: The probability of accepting the null hypothesis when it is false.

0.5 Abbreviations

α: the shape parameter for the Γ distribution

AIC: Akaike Information Criterion

bp: base pairs

DNA: deoxyribonucleic acid

F85: Felsenstein 1985 (model of evolution)

F84: Felsenstein 1984 (model of evolution)

Γ: the gamma distribution for modeling among-site rate variation, where the shape is described by the parameter α

GTR: general time-reversible (model of evolution)

HIV-1: human immunodeficiency virus subtype-1

HKY: Hasegawa-Kishino-Yano (model of evolution)

hLRT: hierarchical likelihood ratio test

I: proportion of invariant sites

i.i.d.: independent and identically distributed

ILD: incongruence length difference

indel: insertion or deletion mutation

JC: Jukes-Cantor (model of evolution)

***k*:** set of all plausible trees used in an SH test

K81: Kimura 1981 (model of evolution)

K3ST: Kimura three substitution type (model of evolution)

***l*:** the branch lengths of a tree; measured in number of substitutions per site

L_{ML}: the likelihood of the ML tree

LnL: negative natural logarithm of the likelihood

$L_{truetop}$: the likelihood of the data when it is constrained to the true topology

$L_{truetree}$: the likelihood of the data when it is constrained to the true tree and the parameters of the true model

M: model of evolution that consists of parameters that specify the relative rates of nucleotide substitution (rate matrix) and the initial base frequencies.

M_0: the null model of evolution, usually user specified

ML: maximum likelihood

mtDNA: mitochondrial DNA

NNI: nearest neighbour interchange

θ: the parameters of the model of evolution, M, and the branch lengths, l

REV: alternative notation for GTR used by Seq-Gen

RNA: ribonucleic acid

rRNA: ribosomal RNA

***S* distribution:** the null distribution for site-pattern variation

***S**:** the test statistic of the oberved data for site-pattern variation

simMODEL$_1$estMODEL$_2$: experimental notation used to indicate the model of evolution used to simulate the sequences and the subsequent model of evolution that was used to estimate trees from those sequences. MODEL$_1$ and MODEL$_2$ can be any of JC, JC+Γ, HKY, HKY+Γ, GTR, GTR+Γ and may also be the same as one another

SPR: sub-tree pruning and regrafting

τ: the topology or evolutionary hypothesis of the taxa in the aligned data set

τ_0: the null topology

τ_1: an alternative topology; usually the topology of an estimated tree

τ_{ML}: the topology of the maximum likelihood tree

T_0: the null tree; consists of topology,τ_0, and branch lengths, l_0

T_1: the alternative tree; usually a refers to an estimated tree

T_{ML}: the maximum likelihood tree

T_{L2}: the estimated tree with the second highest likelihood

TBR: tree bissection and reconnection

tRNA: transfer RNA

UPGMA: unweighted pair-grouping by arithmetic means

0.6 Toolkit

JMP - Statistical package - SAS institute.

MacClade 4.0 - (Maddison and Maddison, 2000)

Modeltest v3.06 PPC- Testing the Model of DNA Substitution (Posada and Crandall, 1998)

PAUP* - Phylogenetic Analysis Using Parsimony* and other methods version 4b10 (Swofford, 1996)

Seq-Gen v1.2.3-1.2.6 - Sequence Generator (Rambaut and Grassly, 1997)

SplitsTree v2.4 – Analysing and Visualising Evolutionary Data (Huson, 1998)

Tree-Edit - Phylogenetic Tree Editor (Rambaut and Charleston, 2001)

TreeView - Tree Viewing Application - (Page, 1996)

0.7 List of Tables

0.8 List of Figures

1 Introduction

1.1 Overview

Evolutionary biology is central to understanding the diversity of life. We use phylogenetic methods to estimate evolutionary histories that cover all manner of life from viruses and bacteria through to plants and animals, which, of course, includes us. Phylogenetic inference requires accurate methods so that we can have confidence in our hypotheses about the relationships among the taxa under investigation. This thesis focuses primarily on the parametric bootstrap, a procedure used to model molecular evolution. Through simulations I examine the accuracy of various phylogenetic methods as we endeavour to answer questions like, "How likely is the answer I got?" or "Are there any other phylogenies that are not statistically different from the answer I got?" The simulated scenarios illustrate the strengths and weaknesses associated with the parametric bootstrap as it is used in topology testing. I examine the accuracy of the parametric bootstrap at recovering site-patterns in empirical data sets and propose its use as a means of model selection.

1.2 Estimating Trees from Molecular Sequences

Since the mid sixties when a number of pioneering individuals (Camin and Sokal, 1965; Edwards and Cavalli-Sforza, 1963; Zuckerkandl and Pauling, 1962) realized how polypeptide sequences may be used to measure the relatedness amongst taxa, the field of molecular systematics has developed rapidly. It was recognised that polypeptides contain a history of past alterations that may be used to estimate the

relatedness of different sequences and it is now common to use both nucleotide and amino acid sequence data to reconstruct evolutionary histories. The evolutionary histories can be estimated from the molecular sequences using models of evolution and the relationships can then be represented as a tree. A tree is the natural representation of the evolutionary history of molecular sequences that are evolving independently. Arguably, there are a few elements that are typically required for reconstructing trees from molecular sequences:

(1) An alignment of the sequences, such that each column represents a homologous site.

(2) An optimality criterion, a method for choosing between hypotheses based on one of parsimony, distance or likelihood (c.f. neighbour-joining which does not require an optimality criterion).

(3) A method of searching for the tree that maximises or minimises the score for the chosen optimality criterion.

(4) A measure or test of statistical support for the result obtained from the tree search.

Unfortunately in evolutionary tree reconstruction, obtaining a tree (step 3) is often seen as the final step of analysis. However, it is of critical importance to continue further and assess the confidence we have in our result (Waddell, 1995). Certain methods, e.g. distance and neighbour-joining, make single point estimates of the topological relationship between the taxa and are thus unable to tell us what other phylogenies might also be acceptable (Felsenstein, 1988). Therefore we are unable to test the results against other hypotheses because sub-optimal results are disregarded

during the optimization process. For these reasons, most authors now advocate a statistical framework that allows us to compare multiple evolutionary hypotheses when estimating trees in phylogenetics.

1.3 Alignment of Molecular Sequences

The alignment of sequences is imperative to the reconstruction of phylogenetic trees. The alignment at any given site is the unit from which a phylogeny is built, such that unaligned or poorly aligned sequences can be misleading or at worse completely uninformative (McCormack and Clewley, 2002). To begin with, homologous sites are aligned, usually with a dynamic programming algorithm implemented in a program like ClustalX (Thompson et al., 1997). In this situation, each homologous site has information on the evolutionary history of those taxa. Issues arise with alignment when sequences contain insertions or deletions (indels) resulting in gaps. These are not easily incorporated into a likelihood calculation (however, see Steel and Hein (2000) for a potential solution) and ignored by parsimony (unless they are treated as a fifth character state). While sequence alignment is not dealt with in this thesis, the value of an accurate alignment cannot be underestimated as the phylogenetic inferences are only as good as the alignment from which they are drawn (McCormack and Clewley, 2002). Great care should be taken to ensure that the alignment generated by the alignment algorithm is checked with visual editing for any anomalous artefacts, as errors at this stage will lead to meaningless results.

1.4 Optimality Criterion

Purely algorithmic methods (e.g. UPGMA and Neighbour-Joining) for tree searching/building have a single step that integrates inference and definition of the

preferred tree into a single step (Swofford et al., 1996). These methods are fast, but have some flaws. One of the most fundamental is that only a single tree is determined. When only a single tree is determined, we have no immediate knowledge about any of the other trees and as a result no strength of support for the tree we have estimated. On the other hand, optimality methods essentially estimate a value or score for each tree, so that a decision can be made as to how good each tree is. The trees are evaluated using either exhaustive, branch-and-bound or heuristic search strategies. These methods come with an increased computational burden, but if one considers empiricism and rigour desirable characteristics in scientific analysis they have several advantages.

(1) There is a score associated with every tree examined.

(2) These scores can be arranged in hierarchy.

(3) Statistical tests can compare the results from the trees to create a confidence set of trees that cannot be rejected for a pre-specified α value, say 0.05.

These advantages can be utilized with a variety of currently-employed optimality criteria. Parsimony-based methods, distance methods and likelihood methods have all been proposed as ways of scoring which trees are the best. Each of these methods chooses to either maximise or minimise a score that assesses the fit between the data and the tree. Distance methods are comparatively quick. They convert DNA sequences into a data matrix and then usually minimise an error of some sort (i.e. least squares). However, the negative aspect of reducing data to pairwise distances is that information is lost in the transformations. An example of this is demonstrated by Penny (1982) where it was shown that several different sets of sequences can yield

the same distance matrix, but given only the distances, it is impossible to go back to the original sequences.

Parsimony methods seek solutions that minimise the amount of evolutionary change required to explain the data under a principle analogous to Ockham's Razor. For example, Camin-Sokal parsimony (Camin and Sokal, 1965) minimises the total number of changes (i.e. the smallest number of changes is the simplest explanation and therefore the most likely). Camin and Sokal applied this method to morphological characters while the same method was also applied to genetic information by Edwards and Cavalli-Sforza (1963). When constructing a tree under this method of parsimony, there are some strong limitations imposed. Firstly, for any given site one of the character states must be specified as ancestral and the other derived. Secondly, this method of parsimony does not allow for the reversal to the ancestral state. These restrictions are relaxed in Wagner parsimony (Eck and Dayhoff, 1966) to provide an implicit model of evolution that is more biologically accurate. Due to the systematic minimisation of evolutionary changes inherent in parsimony-based methods, the true amount of change will always be underestimated because parallel or convergent (homoplastic) changes are not considered (Swofford et al., 1996). The implications for this were brought to the fore by Felsenstein (1978b). He showed that in situations where trees have some long branches the inherent bias in parsimony, due to its lack of consideration for homoplastic changes from the implicit assumption in the model of evolution that mutation events are extremely rare, may lead to problems in phylogenetic estimation, particularly the phenomenon of long branch attraction (LBA). Most substantially it can lead to overconfidence in a possibly wrong topology (i.e. the procedure is "positively misleading"). Although the ideas of parsimony were critical to the development of the field of phylogenetics, the implicit and

undetermined model of evolution as well as the principle of systematic minimisation on which it is based leave us in a position where we should begin to focus on more biologically realistic methods. Many authors consider that previous philosophical arguments that promote the use of parsimony as falsificationist and as a method for assessing degree of corroboration under Popperian philosophy are unjustified (De Quieroz and Poe, 2003). In addition, a relatively recent simulation based study and comprehensive discussion on the relevance of a choice between parsimony and likelihood is presented by Swofford et al. (2001). The authors indicate the strengths and weaknesses of parsimony and leave little room for doubt that statistical methods like maximum likelihood should be chosen when estimating trees from sequences.

Maximum likelihood (ML) was first successfully applied to DNA for phylogenetic studies by Camin and Sokal (1965) then Felsenstein (1973; 1981a). Since then, the strong evidence for LBA and the identification of the poor performance of parsimony based inference (Felsenstein, 1978b), has lead to a surge in the use of likelihood methods. In the Annual Review of Genetics, Felsenstein (1988) wrote:

> *The most effective way of thinking about the inference of phylogenies is to adopt a statistical point of view, as with other kinds of data analysis. It is then seen simply as making an estimate of an unknown quantity, in the presence of uncertainty, and using a probabilistic model of the evolutionary process.*

However, the computational complexity of obtaining maximum likelihood solutions to problems that involve numerous alternative hypotheses has been the limiting factor to the general use of likelihood methods. ML is now implemented as an optimality criterion in many software packages (e.g. PAML, PAUP*, PHYLIP). ML methods perform a sitewise evaluation of an evolutionary hypothesis (topology) in terms of the probability of obtaining a particular distribution of nucleotides across taxa at a certain site given the tree and the model of evolution. By treating the topology of the tree as a

parameter, the likelihood, L, of the parameters (tree and model), θ, across all sites, i from 1 to n given the sequences data, x, is defined by:

$$L = L(\Theta \mid x) = \Pr(x \mid \Theta) = \prod_{i=1}^{n} \Pr(x_n \mid \Theta)$$

where x_n is the distribution of nucleotides across sequences or taxa at a given site.

The likelihood score associated with a tree is usually a very small number, and as such we most often express it as the -ln of the likelihood (-LnL). The use of likelihood has many benefits for the user and is undoubtedly the method with the best-understood statistical basis. One advantage is that the assumptions made in this approach are all explicit and can therefore be checked against real data. A further strength of maximum likelihood methods due to their statistical basis is that they lend themselves to comparisons of different trees, parameters and models, so that competing hypotheses can be tested. The ease of formulating and testing hypotheses is one of the strengths of likelihood (Huelsenbeck and Crandall, 1997). The likelihood axiom states that for any two models, θ_1 and θ_2 being compared, if $L_1 > L_2$ then L_1 is better supported than L_2, and L_1/L_2 measures the strength of that evidence (Hacking, 1965). Statistical estimators are evaluated according to their consistency, efficiency and bias. Consistency is the property of an estimator to converge on the population value as the sample size tends to infinity (Swofford et al., 1996) (i.e. as the number of sites tends to infinity). Efficient estimators that have the property of minimum variance and if unbiased, converge to the true estimate more rapidly (Huelsenbeck and Crandall, 1997). Some types of likelihood, as applied in phylogenetics, have been shown to possess the properties of consistency and efficiency (minimum variance) (Gaut and Lewis, 1995; Rogers, 1997; Yang, 1994b)

and thus provides phylogeneticists with more options in a mathematical framework than other methods. In tandem these properties indicate that maximum likelihood certainly is a potentially powerful procedure. However, it is computational implementation that remains the largest barrier.

1.5 Searching "Tree-Space"

Finding the best tree as judged by the optimality criterion is no small problem. In fact, the number of unrooted binary trees increases super-exponentially (Felsenstein, 1978a). As the number of trees increases, it quickly becomes impossible to search through all possible trees (Figure 1.1). To address the enormous computational burden of exhaustively searching for trees, the branch-and-bound algorithm was applied to tree searching (Hendy and Penny, 1982). Branch-and-bound is still guaranteed to find the tree that maximises or minimises the optimality criterion under which we are searching, but its computational efficiency is data set dependent, and its use is still limited to data sets with no more than approximately 20-25 taxa.

In cases where the data sets are so large that exhaustive and branch-and-bound searches are no longer practical, we usually adopt heuristic searches that sacrifice the guarantee of finding the optimal tree for the chance to obtain a result. As the number of possible topologies is extremely large, the best that can be done in practice is to make small changes to a starting topology until we find one on which we cannot improve (Felsenstein, 1981b). A typical heuristic search will generate such a starting tree through stepwise addition and then follow it up with a series of branch swapping options. The three branch swapping algorithms that are implemented in PAUP* are Nearest Neighbour Interchange (NNI), Sub-tree Pruning and Regrafting (SPR) and Tree Bisection-Reconnection (TBR). Each of these perform different small alterations

to the topology. Although TBR is the most computationally expensive, it has been shown that SPR is equally efficient at reaching the optimum from the given start tree (Takahashi and Nei, 2000). It is clear that this process is not guaranteed to find the ML tree, but it is certainly able to give us a local optimum. The process for estimating the start tree by stepwise addition is known to be order sensitive. We can exploit this by changing the order in which we add sequences and perform the operation multiple times. This also provides opportunity to search as much of the tree–space as possible, and thus increases the probability of finding the global optimum.

Number of Taxa, *n*	Possible Unrooted Bifurcating Trees
3	1
4	3
5	15
6	105
7	945
8	10,395
9	135,135
10	2,027,025
11	34,459,425
12	654,729,075
13	13,749,310,575
14	316,234,143,225
15	7,905,853,580,625
16	213,458,046,676,875
17	6,190,283,353,629,375
18	191,898,783,962,510,625
19	6,332,659,870,762,850,625
20	221,643,095,476,699,771,875

Figure 1.1 The number of bifurcating trees.
It is very clear that due to the rapid increase of the number of possible unrooted bifurcating trees, attempting to search through all of them soon becomes an intractable problem.

1.6 Models of Evolution

We have chosen to use likelihood for the benefits of a statistical approach. A particularly useful aspect of statistics is its ability to attach probabilities to chance outcomes, using a statistical model as the optimality criterion (Waddell, 1995). A number of nucleotide substitution models have been developed to deal with the nucleotide substitution process. The possible substitution processes are outlined in

Figure 1.2. Stochastic models of evolution that can be defined through parameters values for a-f, have essentially two parts:

(1) The initial frequencies of the character states.

(2) The mechanism of the evolution of the characters. This typically includes a relative rates matrix which determines the relative probabilities that a nucleotide, *i*, at site *x*, changes to nucleotide *j*, along a given branch of the tree and is modeled by a time-homogenous Poisson process (Cox and Miller, 1977).

Further parameters including the proportion of invariant sites, I, and a gamma parameter, Γ, to model rate variability across sites are now quite commonly incorporated into a model of evolution as well (Yang, 1996a).

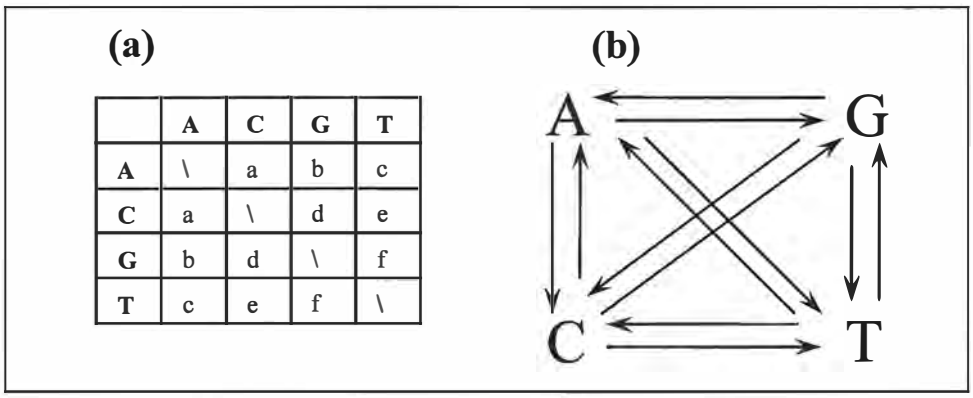

Figure 1.2 Nucleotide substitutions.
(a) The relative rate matrix (r-matrix) is most frequently symmetric and contains values for a-f that describe the relative rates at which the corresponding nucleotide substitutions occur. The GTR model of evolution allows for variation in all of a-f for the various rates of nucleotide substitutions. Simpler models make assumptions that force any of a-f to be equal to one another, thus reducing the number of parameters and simplifying the model. At any given moment, the nucleotide frequencies are assumed to be at equilibrium. (b) The possible nucleotide substitutions between the four bases. Horizontal arrows represent changes that are transitions (purine to purine, A ←→G, and pyrimidine to pyrimidine, C←→T). Diagonal and vertical arrows indicate the possible transversions (purine←→pyrimidine).

Jukes and Cantor (1969) specified the simplest model of evolution. The JC model assumes nucleotides are present in equal frequencies and that the probability that any

nucleotide changes to any other nucleotide during the course of evolution is exactly the same (i.e. all relative rates, a-f, in the r-matrix are equal to 1). This model matched well with the computing power of the day and was certainly a valuable initial model for the analysis of molecular sequences. However, once empirical data are considered, it becomes apparent that this is quite an unrealistic model of evolution. Parameters have slowly been added to the initial JC model often reflecting the need for modification where the JC model had performed poorly. The Kimura-2-parameter model (Kimura, 1980) allowed for a difference in rates of change between transitions and transversions. This was extended to the K3ST model (Kimura, 1981) which allowed a single rate of transitions and two different rates of transversions. Models of evolution were further extended to allow for differences in base frequencies in models such as the F84 (Felsenstein, 1984) and the HKY85 (Hasegawa, Kishino and Yano, 1985). All of these models are essentially assumptions or special cases of the general time-reversible (GTR) model of evolution (Lanave et al., 1984; Tavare, 1986). Within the relative rates matrix (Figure 1.2) of the GTR model we are able to specify the rates of change among all nucleotides. The only limitation on the GTR model is that an A→G is the same as a G→A. A consequence of this time-reversibility is that the tree may be rooted at any position without any change in tree length (Swofford et al. 1996). Leaving a tree unrooted is also an advantage because it allows us to reconstruct trees without worrying about how the likelihood will change depending on the position of the root on the tree. Full asymmetric models are available (i.e. time-non-reversible models where A→G ≠ G→A). These models have increased complexity due to the number of parameters incorporated into the model and at present this complexity is too great to use effectively.

1.7 The Problem with Phylogenetic Estimation

By far the most important problem with phylogenetic estimation is that trees reconstructed from a nucleic acid sample may not lead to the recovery of the true tree, because the amount of data sampled is finite and the true evolutionary process is unknown. This error that contributes to this uncertainty is essentially two-fold. Sampling error is the deviation between the population parameter and the estimate of that parameter from the sample, owing simply to the "random" selection of the sequences used. By definition, if the model is correct and the method of tree reconstruction is consistent, as the sequence length tends to infinity the sampling error tends to zero. Systematic error is an error in the method or in the assumptions of the method; often in problematic cases we more frequently reconstruct the wrong tree and thus increase our confidence in an incorrect solution as the amount of data increases (Rogers, 1997; Yang, 1997). It most frequently arises through a critical violation of the assumptions of our data or because the methods are flawed under certain conditions (e.g. LBA). And we will always have to make assumptions; Felsenstein wrote in the Annual Review of Genetics (1988):

> *It is clear from the failings of different methods in particular cases that they all have assumptions; no method allows one to make inferences about evolutionary patterns in a well-justified way without making any assumptions about evolutionary processes.*

More recently, in response to recent advances in modeling evolution Sullivan and Swofford (2001) wrote:

> *Nevertheless, even with these improvements, there is no reason to believe that even the most general and parameter-rich evolutionary models currently available capture all the nuances of the processes that have generated any particular set of sequences.*

For example, it is always assumed that sites are independent and identically distributed (i.i.d.) (Cavender, 1978). Independence implies that characters do not influence each other in any way, and identically distributed implies that the same underlying principles influence each site in the same way. It is unlikely that any sequences will fully satisfy these assumptions, but non-i.i.d. models are currently intractable. However, our methods may be robust to the assumptions such that we can gain considerable mileage even with essentially "incorrect" models. Box (1976) said "*All models are wrong, but some are useful*". In fact, maximum likelihood (and some other methods) appears to be robust to violations of many assumptions (Huelsenbeck, 1995; Huelsenbeck and Nielsen, 1997; Schoniger and Von Haeseler, 1995). Therefore, even in the absence of complete knowledge, we can use these hypothetical models of the evolutionary process to derive (or otherwise justify) tree inference methods that would be free of systematic error, *if the assumed model were correct* (Swofford et al., 1996). In other words, a model does not have to be perfect in order for it to be useful.

Different trees constructed using likelihood-based approaches give rise to T_{ML}, T_{L2}, T_{L3} etc. (where T_{ML} is the maximum-likelihood tree and T_{L2} is the tree with the second highest likelihood etc.). The maximum likelihood tree is the tree with the highest likelihood, or the lowest -LnL (as it is usually represented). For any given data, the difference in log-likelihood scores between LnL_{ML}, LnL_{L2}, LnL_{L3} etc. comes from two sources. The first source is in the variation of the branch lengths and the second is in the tree shape or topology. Together the topology and the branch lengths provide the parameters for the tree. The more critical of the two sources of log likelihood difference is arguably that obtained through a change in topology. This is principally because the topology of the tree represents an evolutionary hypothesis.

Therefore, an alteration to the topology of a tree is also an alteration to our interpretation of the evolutionary history of the taxa on the tree. For this reason, it is of critical importance to statistically test if there are multiple trees that are supported by the observed data.

1.8 The Necessity of Statistical Testing

One way of thinking about statistics is that it is the science of decision making under uncertainty (Chernoff and Moses, 1959). It involves finding a solution by using a procedure to choose an action. In some cases, the action may be a choice between two ideas (hypotheses). We can gather data and use this as evidence to favour one hypothesis over another. However, in some cases the data may not favour one over another and instead we are unable to make a decision as both hypotheses appear to be equally good explanations of the data. If we do not test the hypotheses statistically to ascertain if one is favoured over the other, we have not completed the analysis, and cannot interpret the results with any degree of confidence leaving the result incomplete and potentially misinformative.

The necessity of statistical testing of topologies can be no better understood than through an example. Wainright et al. (1993) produced a tree topology (Figure 1.3) that grouped the fungi more closely with the animals than the plants. Until this study, it had been more commonly thought that the fungi and the plants would have a common ancestor not shared with animals. The claim was made that there was support for such a tree using the phylogenetic bootstrap. However, many of the bootstrap proportions were, in fact, rather low.

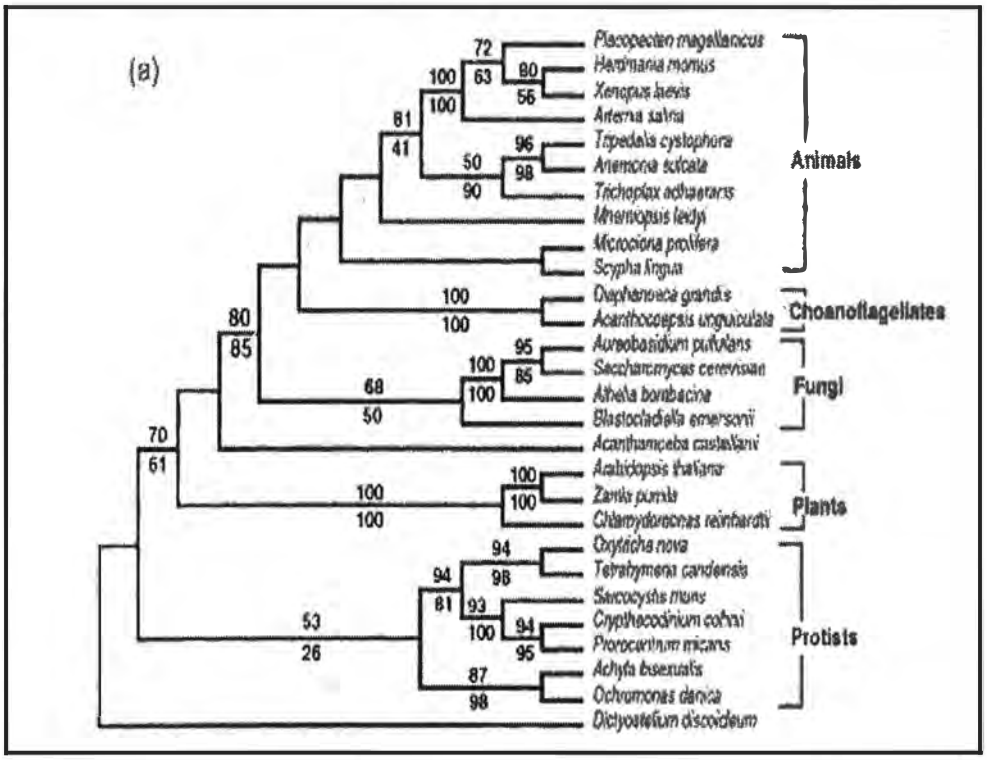

Figure 1.3 The phylogenetic tree published by Wainright et al. (1993).
We observe that in this topology the animals and the choanoflagellates share a common ancestor with the fungi before sharing a common ancestor that unites the fungi with the plants. The numbers above each branch indicate the percentage of bootstrapped maximum likelihood trees that support that clade. The numbers below each branch indicate the percentage of bootstrapped neighbour-joining trees that support that clade.

Rodrigo et al. (1993) used the Kishino-Hasegawa test (Kishino and Hasegawa, 1989) to test whether or not a topologically distinct tree, that grouped the plants with the fungi before a common ancestor with the animals (Figure 1.4), would be statistically different from the tree presented by Wainright and co-workers. Here the KH test, a test of topologies, was correctly used as both topologies were specified *a priori* and the results showed that for those data, the two topologies were, in fact, not statistically different from one another. Furthermore, it still remains a question as to how many other topologies would also not be significantly different from the tree proposed by Wainright and co-workers. This is a task for another test, the Shimodaira-Hasegawa test (Shimodaira and Hasegawa, 1999), which allows for multiple comparisons. This example illustrates the need to consider more than just the

ML tree and that a final step of evaluating an estimated topology within a statistical framework is pivotal to biological inference.

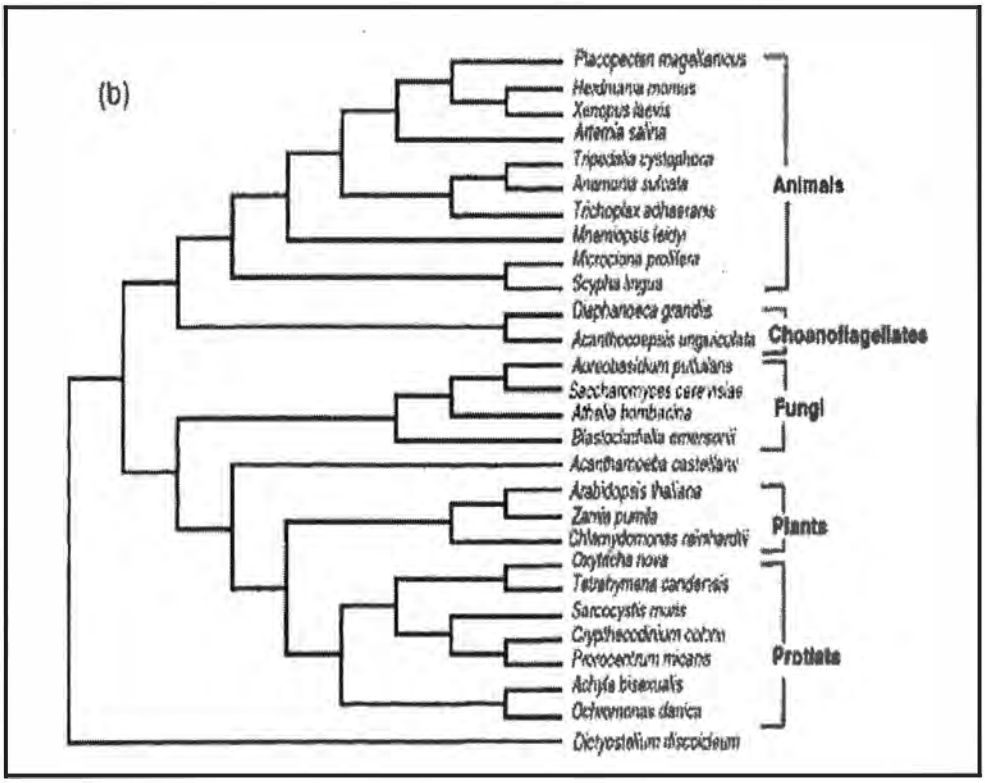

Figure 1.4 The alternative phylogeny proposed by Rodrigo et al. (1993).
The outgroup now joins the tree at a node between the animals and the fungi. Apart from this change, all other branching patterns were maintained. The effect on the evolutionary hypothesis is that the fungi now share a common ancestor with the plants and protists before joining with the animals and the choanoflagellates.

1.9 Topology Testing in Molecular Phylogenetics

There is no current generally applicable method capable of assessing the phylogenetic validity of the molecular data that investigators are generating. This is in part due to the current level of understanding of tests of topologies, which is insufficient. Felsenstein promoted the use of the non-parametric bootstrap to generate a level of confidence for each clade (Felsenstein, 1985). This uses a sampling-with-replacement strategy of the aligned sites to generate pseudoreplicate data sets with the same number of taxa and sites. Each site from the original data set may appear 0, 1, or

many times. Any number of replicate data sets can be generated, but typically, 100 to a 1000 are created. A tree is estimated for each data set using the same method as was used on the original data. The number of times the replicate data sets support a clade within the tree, is a measure of support for that clade. A single consensus tree that contains the clades with the highest bootstrap support is usually used to summarise the bootstrap pseudoreplicate trees. Due to accusations of bias in the non-random nature of the character sample space and this relatively simplistic application of the bootstrap we continue to strive for better methods of assessing the confidence in our tree (Sanderson, 1995). Corrections to the standard application of the bootstrap can be made using more elaborate methods (Efron, Halloran and Holmes, 1996; Rodrigo, 1993; Zharkikh and Li, 1995), but as expected, this comes at the expense of considerably more computation time.

There are a number of tests available that examine the level of confidence that one might have in a reconstructed phylogeny estimated from molecular sequences. The majority of these techniques have been designed for specific uses, and there is currently no general test that can be applied to the number of situations that regularly occur in phylogenetics. Inappropriate use of a test (i.e. when not used for its designed purpose) can lead to the violation of a number of critical arguments in the derivations. This only serves to provide misleading results, often in the form of bias and inaccurate setting of significance levels. In many cases, the result is overconfidence in the wrong tree, which is not an insignificant problem. The first test of topologies was a non-parametric test developed for parsimony data (Templeton, 1983). The Templeton test utilizes a Wilcoxon ranked sums test of the relative number of steps required by each character on each of the respective trees in the comparison. This simple test compares the number of characters that favour each of the two trees and

test the results against a binomial distribution. For likelihood-based tree estimation, two tests of topology (the KH test and the SH test) are currently implemented in some of the phylogenetic software packages that are available (e.g. PAUP*). A third test, the Swofford-Olsen-Waddell-Hillis test (SOWH test), requires the combination of at least two software packages with guidelines available online at http://www.ebi.ac.uk/goldman/test/SOWHinstr.html. These three tests are in principle equally relevant to nucleotide and amino acid data.

A situation for wanting to use topology tests arises when we consider topological hypotheses 'A' and 'B'. If we specify these competing hypotheses (topologies) before starting our analysis and we obtain hypothesis 'A' when we expect B, we can correctly use the KH test since both A and B are specified *a priori*. Typically the problem is more unpredictable than this and we frequently obtain a third hypothesis 'C'. In this situation it is incorrect to use the KH test as 'C' is not an *a priori* hypothesis for this analysis. However, we would still be interested in comparing it to 'A' and 'B' to show whether or not 'C' is preferred to the exclusion of 'A' and/or 'B'. Alternatively if 'A' and 'B' are not significantly different from 'C' then we should consider all of 'A', 'B' and 'C' to be members of the confidence set of topologies all of which are not significantly different in their ability to explain the data. Suffice to say that when we get a "surprising" result (i.e. one that we did not specify *a priori*) we are interested in assessing the confidence in our result using a statistical test.

1.9.1 The Kishino-Hasegawa Test

The Kishino-Hasegawa test (KH test) was developed over a decade ago (Kishino and Hasegawa, 1989) to test between competing evolutionary hypotheses (tree topologies). The KH test has the ability to be used on both parsimony and likelihood

data by estimating the standard error and the subsequent confidence intervals for the difference in log-likelihoods between two topologically distinct phylogenetic trees (Goldman, Anderson and Rodrigo, 2000). The test was designed for, and initially applied in, situations where the trees being compared were specified *a priori*. Under these conditions the comparison of the trees is independent of the sample at hand and therefore the test is being correctly used (Swofford et al., 1996). Software packages that currently implement the KH test cannot discern between users that have specified their hypotheses *a priori* and those that have not. As a result the problem is magnified, by users who do not understand the correct application of the KH test.

Unfortunately, the most typical situations are those most likely to fall foul to misuse of the KH test. The difficulty arises when we consider that for a given data set, if we estimate the maximum-likelihood tree (T_{ML}) and then wish to compare it to the trees with the second, third, or fourth highest likelihoods or a previously hypothesised tree, T_0, we violate one of the most fundamental assumptions of the KH test. In this situation, the hypothesis under investigation actually changes. Following the notation of Goldman et al. (2000), for two trees specified *a priori* one would have a null hypothesis:

$$H_0: E[\delta] = 0$$

where our test statistic, δ, is the difference in log-likelihoods of the two trees being compared. This is equivalent to the result of a likelihood ratio test. For any one data set, stochasticity and sampling would ensure that the value of δ is unlikely to equal zero, but for a large number of data sets the expectation for the test statistic, δ, is that the difference in log-likelihoods would be zero. However, if one of the trees is

the ML tree, deduced *a posteriori* then $L_{ML} - L_1$ will always be greater than zero, because the maximum likelihood tree will always have the higher likelihood (i.e. the maximum likelihood). In these circumstances, when it is assured that δ will always be greater than zero, the null hypothesis becomes:

$$H_0: E[\delta] > 0$$

The design of the test relies so fundamentally on the independent selection of the topologies that many of the arguments in the derivations cannot be upheld if this is not the case. Two sided testing is also no longer appropriate in this situation as we are expecting deviation only in one direction. Intrinsically none of these are flaws with the KH test, but incorrect application of it would clearly lead to positively misleading results and, specifically, overconfidence in an incorrect tree.

Due to the incorrect reconstruction of estimated trees as discussed earlier, it is important to provide a confidence set for the trees (those trees that we are unable to reject at a pre-specified α level, say 0.05). This confidence set may contain the number of other topologies not significantly different from the ML tree. Of course, this is not to say that such a difference in topology is not biologically significant. Each topology represents a different hypothesis of the relationship between the taxa under investigation. Therefore any difference noted in topology actually represents a different evolutionary hypothesis for the relationship between those taxa. Such discrepancies are by no means small, and it is therefore important that we use appropriate tests that enable us to determine the reliability of published data. Therefore it is extremely important to look at more than a single tree with only bootstrap proportions assigned to the internal nodes. To draw biological conclusions

(our ultimate goal) we need to be certain that our results are significant statistically and then biologically.

1.9.2 The Shimodaira-Hasegawa Test

The Shimodaira-Hasegawa test (SH test) was developed as a modification to the KH test to allow for a correct application to multiple topological comparisons and partly alleviate some of the associated *a priori* selection bias. The SH test was specifically devised due to incorrect application of the KH test in a multiple topological comparisons framework, which leads to overconfidence in the wrong tree (Shimodaira and Hasegawa, 1999). In this test, all trees that the investigator believes could possibly explain the data must be considered if we are to find the true topology (Goldman, Anderson and Rodrigo, 2000). The set of all plausible trees, k, must once again be specified *a priori* to ensure that significance levels are accurate. To ensure that k does contain the true topology we may need to be conservative and choose all possible trees for the given number of taxa. However, as the results of the SH test are dependent on the number of topologies in k, it may be important to limit the size of k, through the prudent application of prior knowledge (Buckley, 2002; Goldman, Anderson and Rodrigo, 2000; Strimmer and Rambaut, 2002). If a large set of possible topologies cannot be avoided, this provides another significant area of computational intensity.

The null and alternative hypotheses for the SH test are:

H_0: all $T_x \in k$ are equally good explanations of the data

H_A: some or all $T_x \in k$ are not equally good explanations of the data

The SH test generates non-parametric bootstrap replicates for each of the topologies. Each topology is then tested to see whether it falls within the 95% confidence interval for $E[\delta_x]$, the expected difference in log likelihoods between the two topologies. A one sided test is appropriate as we know that $L_{ML}-L_x > 0$. The observed problems with the SH test are that although the topologies are visibly different, these are not often noted as being statistically different. This shows that the SH test is conservative (Buckley, 2002; Goldman, Anderson and Rodrigo, 2000; Shimodaira and Hasegawa, 1999) or less likely to reject the null hypotheses when they are false, when compared to other methods. This illustrates a tendency to include unlikely trees in the confidence set (Strimmer and Rambaut, 2002). This property of the SH test makes it largely unsuitable for inferring confidence intervals for tree topologies especially since the extent of its conservative nature is still unclear.

1.10 Monte Carlo Simulation in Topology Testing: The SOWH Test

The theory of statistical testing using a parametric or Monte Carlo based approach involves the simulation of replicate data sets that precisely conform to the assumptions of the null hypothesis. These replicates are guaranteed to be drawn from the distribution induced by the null hypothesis and their distribution therefore is a parametric estimate of the null hypothesis distribution of that statistic (Efron, 1985; Goldman, Anderson and Rodrigo, 2000; Huelsenbeck, Hillis and Nielsen, 1996). If the distribution of possible values for the statistic is known, it is possible to distinguish between acceptable and unacceptable deviation from expectations (Goldman, 1993). In addition, Felsenstein (1988) wrote:

> *One method, the parametric bootstrap, consists simply of taking the best estimate of the tree, simulating new data sets of the same size by evolution occurring along that tree under the postulated model and then using the variability among estimated trees from those simulated data sets to assess how much variability there was in the original estimate. This is one of the best uses of simulation, and should be done more frequently.*

It is possible from the distribution to create a parametric bootstrap LRT to assess whether an *a priori* selected topology τ_0 is supported by a sequence data set or should be rejected in favour of another topology. The SOWH test (Swofford, Olsen, Waddell and Hillis Test) was named by Goldman et al. (2000) after the authors who first suggested its application to phylogenetic tests of topologies (Swofford et al., 1996). The SOWH test endeavours to evaluate the expected topology, τ_0, against an alternative, the topology of T_{ML}, where T_{ML} does not have the same topology as τ_0. The null hypothesis that τ_0 is the true topology is based on prior knowledge that guides our expectation.

H_0: τ_0 is the true topology

H_A: some other topology is true

The SOWH test for testing these hypotheses is outlined below:

(1) Calculate the test statistic, δ, from the real data where $\delta = LnL_{ML} - LnL_0$. LnL_0 is the log likelihood of the data constrained to the topology τ_0. (Note: the standard likelihood ratio test (LRT) statistic is equal to $2(LnL_ML - LnL_0)$, but since it is only a multiplier of 2 this does not affect the outcome of the Monte Carlo simulation results).

(2) Use the model of evolution estimated from the data constrained to τ_0 to simulate the parametric replicate data sets, *i*.

(3) For each data set, i, estimate T_{ML} and the likelihood of the data constrained to τ_0 to create Δ_i values, which collectively form the Δ distribution.

(4) Finally, test the value for δ, obtained from the original data set against the distribution of Δ_is obtained from the parametric replicates.

Using an α level of 5%, if our observed value is atypical of the parametric distribution and therefore cannot be accounted for by chance alone, we can obtain a measure of significance against the null hypothesis that our topology is not the true topology. An attractive feature of the SOWH test is the increased power associated with a parametric test, because it can use the form of the distribution that gives rise to the data (this is a feature not available in non-parametric tests (Goldman, Anderson and Rodrigo, 2000). The cost associated with the benefit of this increased power in the statistic is an increase in the reliance on the assumed model, so one must only consider using a robust model – a rather contentious assumption. Typically the use of parametric bootstrap approaches in other areas of phylogenetics have proven to be powerful and robust to deviations from the assumed model so long as an attempt has been made to use a model sufficiently complex to encompass the major features of the data (Goldman, 1993; Hillis, Mable and Moritz, 1996; Huelsenbeck and Crandall, 1997; Yang, Goldman and Friday, 1994)

The dependence on the model can be directly attributed to the fact that each replicate is simulated under the same model as was used to generate the sequences. The resulting difference in log likelihoods will then be very small leading to a very tight distribution. The other major documented difficulty with the SOWH test is type I

error (Buckley, 2002). This is to say that we reject our null hypothesis when it is true. This is likely to be partly explained through the model misspecification, but in addition, I believe that this may be attributed to the test not being a test of topologies, but a test of trees. Evidence that supports this suggestion is presented as part of chapter 2.

1.11 Simulation Studies

A fundamental paradigm in phylogenetics is the use of biological data to estimate the evolutionary history of the taxa, without ever being able to guarantee that the truth has been obtained. Methods that unequivocally estimate the same result may not necessarily be correct, having elements of bias that cause this to be the case. To assess the applicability of methods, we create the situation in which the truth is known about the data. Monte Carlo methods generate data from a model that can be specified by the user. If the null hypothesis is fully specified, then the probability distribution for the data is known. The null hypothesis in phylogenetic simulations is specified to include a topology, τ_0, the branch lengths, l_0, and the model of evolution, M_0. We simulate according to the parameters of the null hypothesis to generate sequences that conform to the distribution of that null hypothesis.

By knowing the truth about the way these sequences were evolved, we can make comparisons between the null hypothesis and each of the data sets that have been simulated. Each of the replicate data sets are equivalent to a typical biologically obtained data set (i.e. an alignment of homologous sequences). In biologically obtained molecular sequences the null hypothesis is unknown and we have only a few genes where the phylogenetic signal in the molecular sequences is sufficient to reconstruct the evolutionary histories. Each of these genes, while they may have the

same true evolutionary history, will almost certainly have different selection pressures (Li, 1997), and therefore different models of evolution for the same evolutionary history. In contrast, the advantage of the simulation study is that each data set conforms to the same null hypothesis and therefore, the combination of the results from individual replicate data sets provides us with a distribution for the result and not a point estimate. In chapters 2 and 3 of this thesis, trees estimated from sequences are compared back to the original null hypothesis tree to see if the topologies are the same. The SOWH test is analysed for type I error which can be calculated from each instance in which the null hypothesis is rejected, because we know that it is true (and therefore should not be rejected). In chapter 4, replicate data sets are simulated under models estimated from real data. The variation in the site-patterns between the replicates and the original is used to compare models and investigate evolutionary patterns.

2 Comparing the 95% Significance Level of the KH and SOWH Tests

2.1 Outline

Now that I have discussed the KH and SOWH tests, I will move on to their implementation in a testable framework. Intuitively, two equally valid, but different approaches to analysing the same data should produce the same answer (Strimmer and Rambaut, 2002). So it is not insignificant if the two different approaches do not agree. However, a direct comparison between the results of these two tests is not straightforward. This is because the tests are actually designed to examine different null hypotheses. The KH test evaluates whether two topologies, τ_1 and τ_0, are equally good explanations of the data where the expected difference in log likelihoods is zero (H_0: $E[\delta] = 0$). The SOWH test evaluates whether or not τ_1 is the true topology by determining whether the difference in log likelihoods is typical or atypical of the null distribution. So, how do we compare statistical tests with different null hypotheses? Classical Neyman-Pearson theory uses the data as a measure of evidence and here we compare the 95% significance levels of the two tests using the simulated data as our evidence.

The first standard requirement for a simulation study such as this is to fully define the null hypothesis. Knowing the null hypothesis puts us in a position of certainty when comparing our replicate data sets back to the original. To create the typical biological situation in which we would use tests of topology, we require data sets that although they have been simulated for a given null hypothesis actually do not re-estimate the

null hypothesis topology. The KH and SOWH tests are applied on each of the data sets that do not re-estimate the null topology, and the 95% significance levels of each are calculated, so that they can be compared. For the SOWH test, I have followed the ideas of Swofford et al. (1996) with the method and notation as described by Goldman et al. (2000) and as implemented by Buckley (2002). The KH test is implemented in PAUP* with the exact specifications supplied in the methods section. In this chapter I compare the significance levels (at $\alpha = 0.05$) for the two tests. Therefore, to perform this comparison of the 95% significance level one requires:

(1) A start tree, T_0 that has topology and branch lengths defined by the user.

(2) A model of evolution, M_0 that includes user defined base frequencies, rates of substitution (R-matrix) and the optional inclusion of a gamma parameter, Γ, to model the variable sites.

(3) A program that uses Monte Carlo simulation to parametrically simulate the replicate data sets (aligned molecular sequences) given T_0 and M_0 (Seq-Gen).

(4) A program to estimate the trees from those replicate data sets using maximum likelihood (PAUP*).

(5) An automated procedure for determining the difference in log likelihoods, Δ_i, between the constrained and unconstrained trees for all trees estimated from the replicate data sets, i, that have a topology that is different from the topology of T_0 (my own Perl scripts).

The individual values of each Δ_i from each replicate data set are used to construct the Δ distribution. We are interested in the 95% significance level ($\alpha = 0.05$) of this

distribution and how it compares to the $\alpha = 0.05$ of the KH test. The procedure for generating a Δ distribution is summarised simply in the flow diagram (Figure 2.1).

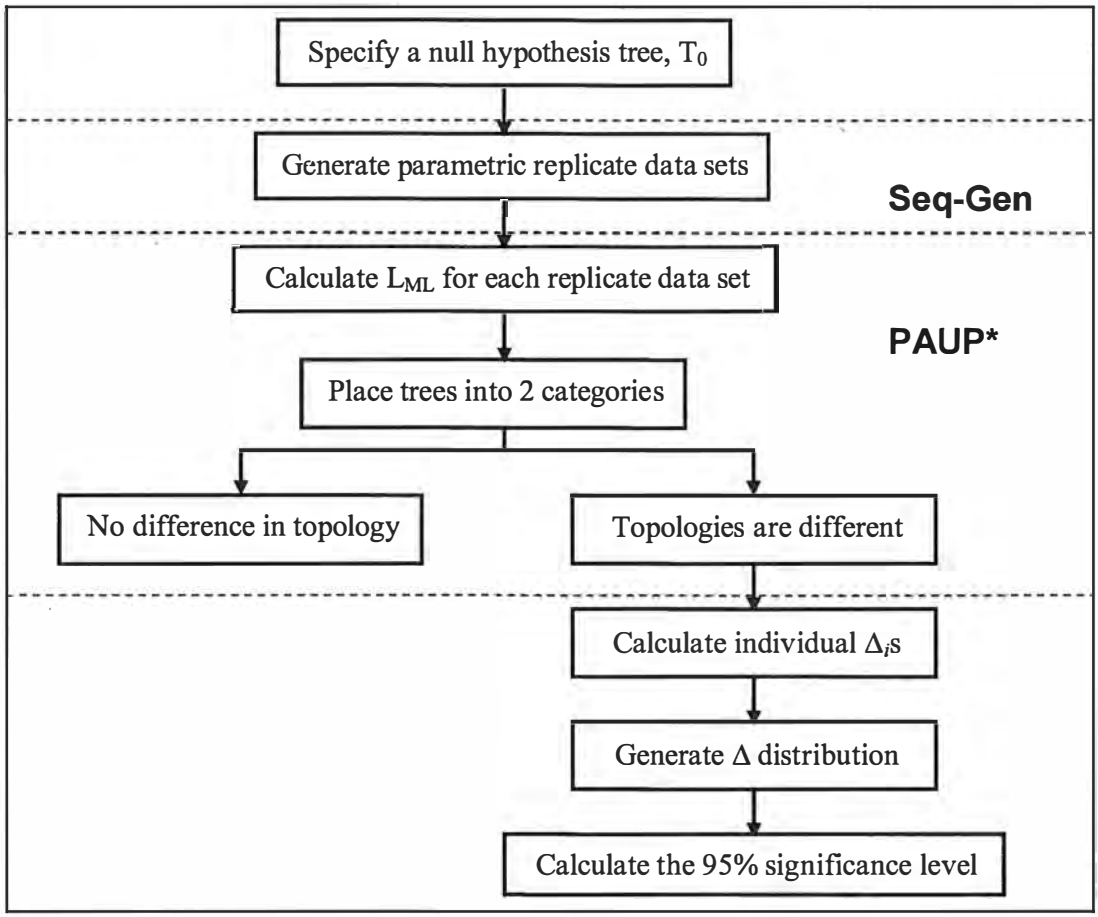

Figure 2.1 A flow diagram outlining the SOWH test.
Two programs Seq-Gen and PAUP* are required to perform the SOWH test. Following their use, I generated the Δ distribution (null distribution) from the likelihood scores of the parametric replicates using Perl scripts

The KH test has an extra step of computation, but fits in conveniently after what has already been performed. The KH test is also only performed on those trees that are not reconstructed correctly (i.e. when the ML tree does not have the same topology as T_0). The KH test is implemented in PAUP* and tests whether or not the difference in likelihood between two competing topologies is statistically significant. The following flow diagram (Figure 2.2) shows the main procedures in the method for testing significance on the KH tests. The procedure for both the KH and SOWH tests

is repeated on sequences simulated according to each of six models of evolution corresponding to JC, JC+Γ, HKY, HKY+Γ, GTR and GTR+Γ. Following the simulation of the replicate data sets, each replicate was estimated with each of the six different models of evolution. To bridge the gaps between many of the steps, and to set up the procedures so that they were automated, a few Perl scripts[1] were used.

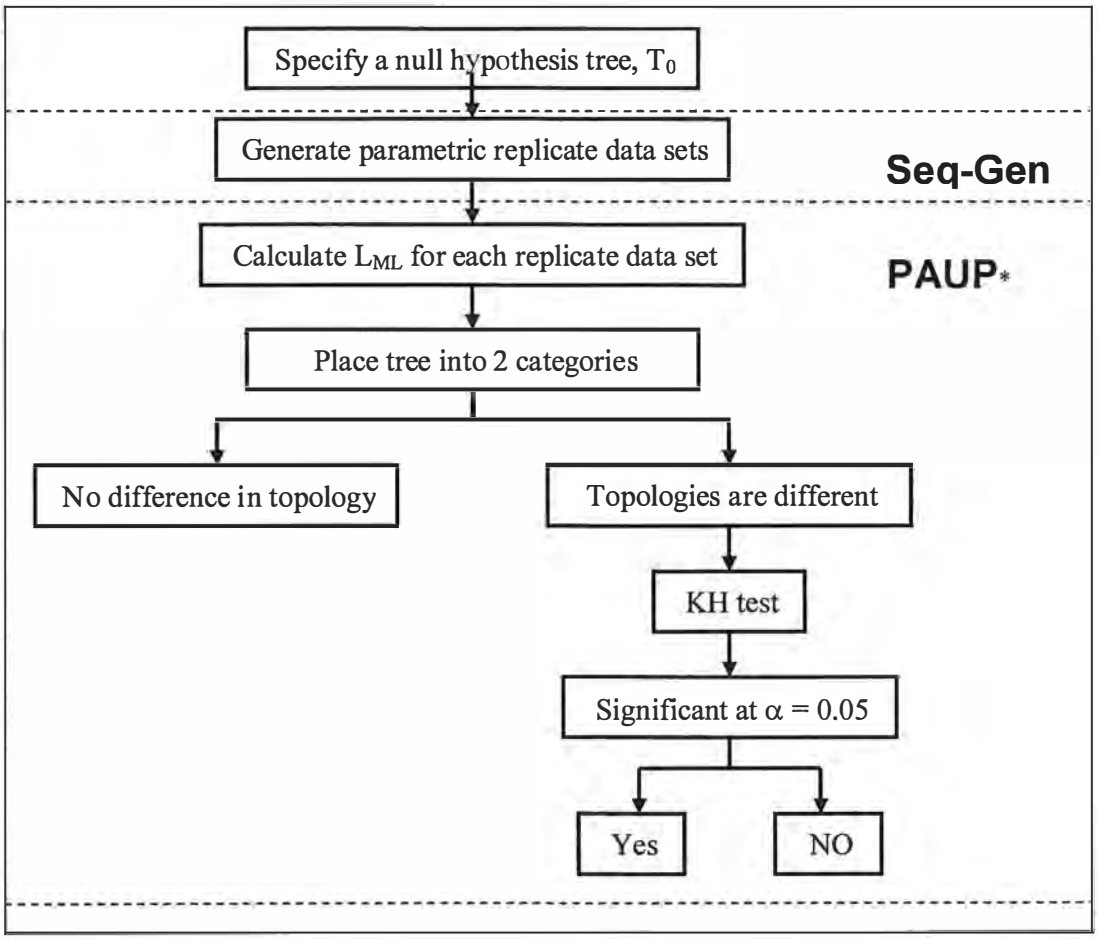

Figure 2.2 A flow diagram outlining the KH test.
The simulation of the replicate data sets requires Seq-Gen. Unlike the SOWH test, the KH test is implemented in PAUP* and can be performed directly on any two topologies when calculating the likelihood scores for each of the topologies.

[1] The function of each script is described in the text, important features are available in the appendix and the full working code is available on the accompanying CD.

2.2 Methods

2.2.1 Choosing a Topology

Traditionally simulation studies in phylogenetics have used four taxon trees (Cummings, 2003; Felsenstein, 1978b; Gaut and Lewis, 1995; Huelsenbeck, 1995; Siddall, 1998). It has been argued that they possess all the required properties of other trees without the difficulty of extracting signal from noise due to large and unknown interaction effects. Some studies have drawn away from using just four taxa (Guindon and Gascuel, 2003; Kuhner and Felsenstein, 1994; Sullivan and Swofford, 2001). For this chapter, I chose to move away from four taxa and used an eight taxon tree as the true tree such that by chance alone we will obtain the true topology with a probability of 9.62×10^{-5}. The consequence of using 8 taxa instead of 4 is a rather considerable increase in the computational burden, but this also allows us to assess the applicability of the methods in a more biologically realistic context, since it is very uncommon to have data sets of size four. The chosen topology for the 8 taxon tree was generated randomly using the random tree function from MacClade (Maddison and Maddison, 2000) and half the terminal branch lengths were intentionally specified long and the other half short (Figure 2.3). The internal branch lengths were specified at the same length as the short terminal branch lengths.

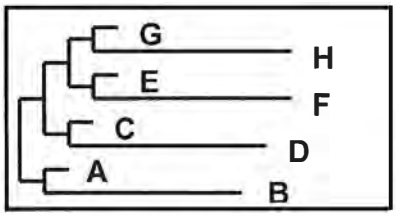

Figure 2.3 The unrooted 8 taxon tree used to simulate the replicate data sets in chapter 2. In this tree, the short branch lengths (all the internal branch lengths as well as those labelled G, E, C and A) are all equal to one another but different from the long branch lengths which are also all equal to one another. The tree, T_1, that was used for the simulations has short branch lengths set at 0.02 while the long branch lengths were set at 0.4 (measured in substitutions per site).

2.2.2 Choosing the Tree Parameters

To investigate topology tests, we must ensure that at least some of the reconstructed topologies of parametrically simulated alignments are incorrectly reconstructed. For the SOWH test and the KH test we require incorrectly reconstructed topologies from our parametric simulation because it is only in the instances of finding a tree that has a topology different from τ_0 that we are interested in performing these topology tests. To achieve a sufficient proportion of incorrectly reconstructed trees, I performed a number of experiments under a trial and error process varying the branch lengths of the input tree, and the length of sequence of the output under a Jukes-Cantor model of evolution. Seq-Gen (Rambaut and Grassly, 1997) is a program capable of simulating sequences from a starting tree under a model of evolution. Using parametric or Monte Carlo simulation, Seq-Gen simulates sequences that, for a given model of evolution and the null tree (topology and branch lengths) should give the correct tree (if the model is correctly specified and we have infinite data). However, if we introduce a systematic error like long-branch attraction (LBA) then we see occasionally that a tree estimated from the simulated sequences will be incorrect.

Throughout the trial and error process it was very difficult to get a substantial number of incorrectly reconstructed topologies using the Jukes-Cantor model of evolution to simulate the sequences of 500 bp in length and then estimate the trees from those sequences. In fact, the branch lengths that were finally used (0.02 for the short branches and 0.4 for the long branches, Figure 2.3) are typical of a "diabolical" tree. Diabolical trees are such that if obtained from real data, would be treated with a great deal of skepticism. The effects of LBA, given the relatively short nature of the sequences, are well documented (Felsenstein, 1978b) and as such would provide

significant concern to investigators using even the best available models. This tree however, was appropriate for generating a reasonable proportion of incorrectly reconstructed topologies for the subsequent assessment of the 95% significance levels of the SOWH and KH tests.

2.2.3 Simulating the SOWH Replicate Data Sets

After the preliminary trial and error experiments used to obtain an appropriate start tree as the null hypothesis, I simulated 1000 SOWH replicate data sets each with eight sequences using the null hypothesis tree in Seq-Gen under models of evolution that corresponded to JC, JC+Γ, HKY, HKY+Γ, GTR and GTR+Γ (see chapter 1 Models of evolution for a description). The command lines used to simulate the sequences under the appropriate models use a number of defaults and are shown in Table 2.1. In the more complex models, the parameters for nucleotide frequencies (HKY and GTR) and the rate matrix (GTR only) are empirical estimates for the HIV-1 *env* gene.

Model of Evolution	Seq-Gen Command
JC	-l 500 -on -n 1000
JC+Γ	-l 500 -on -n 1000 -a 0.4
HKY	-l 500 -on -n 1000 -m HKY -t 2 -f 0.4 0.15 0.21 0.24
HKY+Γ	-l 500 -on -n 1000 -m HKY -t 2 -f 0.4 0.15 0.21 0.24 -a 0.4
GTR	-l 500 -on -n 1000 -m REV -f 0.4 0.15 0.21 0.24 -r 0.018 0.17 0.52 1.22 3.08 1
GTR+Γ	-l 500 -on -n 1000 -m REV -f 0.4 0.15 0.21 0.24 -r 0.018 0.17 0.52 1.22 3.08 1 -a 0.4

Table 2.1 The Seq-Gen commands used to simulate sequences under all models of evolution in chapter 2.
JC is the Jukes-Cantor model proposed by Jukes and Cantor (Jukes and Cantor, 1969), HKY is the model presented by Hasegawa, Kishino and Yano (Hasegawa, Kishino and Yano, 1985) and GTR is the General Time-Reversible model proposed by Lanave et al. (Lanave et al., 1984). The gamma parameter, Γ, was introduced to allow variable rates across sites (Yang, 1994a).

The default settings of Seq-Gen match a Jukes-Cantor (JC) model of evolution with no gamma parameter (i.e. equal base frequencies, equal rates of substitution and an equal rate of among-site variation). -l is the sequence length parameter, which in this

chapter was set to 500. -on generates an output file in nexus format, the standard input format for PAUP*. -n specifies the number of replicate data sets, which is set to 1000. To change the model to allow for more complicated nucleotide substitutions we use -m and specify either HKY or REV[1]. When the model is altered, there are now further options available. -t specifies the ratio of transitions to transversions in the HKY model and was set to 2 for all HKY simulation in this chapter. -r allows us to specify the values for the relative rates matrix for the general time-reversible model. For all simulations in this chapter involving the GTR model of evolution the rate matrix was set at a=0.018, b=0.17, c=0.52, d=1.22, e=3.08, f=1. In general the parameters of the matrix are scaled to have f=1, such that the computational burden is minimised and therefore its inclusion is unnecessary, but Seq-Gen requires its inclusion to run. Most empirical estimates of the gamma distribution shape parameter, Γ, fall in the range of 0.1-0.5 (Yang, 1996a). So, whenever Γ was included in the model of evolution to model the rate heterogeneity, it was set using the command line -a to an α value of 0.4.

2.2.4 Estimating Trees from the Replicate Data Sets

To begin with, I ran both the exhaustive search and a search using the branch-and-bound algorithm to search for the optimal tree. Both methods guarantee that the ML tree has been found, but the efficiency of the branch-and-bound algorithm is data set dependent. In these early searches, branch-and-bound was approximately three-fold quicker than the exhaustive approach, so branch-and-bound was preferred for all further analyses. The likelihood settings of the search depended on which model of

[1] REV is the Seq-Gen notation for the General Time-Reversible (GTR) model of evolution.

evolution was being used for the estimation. Sample nexus files for each are shown in the appendix (Section 7). Each PAUP command block was added using a Perl script (PAUPcommanderMODEL[1]) that attached the appropriate commands after each data set. A total of 36 data groups were created through the simulation and estimation procedure that correspond to which models were used to simulate and estimate the data. For example simJCestGTR is the data group for which all data sets were simulated using JC and estimated using GTR. A limitation encountered was the length of time required to estimate under GTR+Γ. Each simulation took approximately 48 hours and it had been planned that there would be 6000. The computational burden was too extreme for the resources available and these estimations were abandoned.

2.2.5 KH Testing the Replicate Data Sets

To create a more automated procedure for such a large number of simulations, every data set was immediately KH tested after estimation regardless of the whether or not the topology was different from τ_0 or not. Usually, a KH test would only be performed on trees that have a topology different from τ_0, but if two identical topologies are compared, the ΔLnL is equal to zero and the result of the KH test is not significant. The results of the KH tests are logged to the same file as the estimation outputs from the PAUP* display buffer. The results of the KH test are extracted, the

[1] MODEL is specific to the model of estimation and is one of JC, JC+Γ, HKY, HKY+Γ, GTR, GTR+Γ. A copy of PAUPcommanderGTR is on the accompanying CD.

number of significant results tallied and the Δ_i values extracted in a tab-delimited format that can be imported into Microsoft Excel by the Perl script LNLandSIG[1].

2.2.6 Comparing the KH and SOWH Tests the 95% Significance Level

To show how the KH test and the SOWH test compare in terms of the difference in log likelihoods, Δ, at the significance level of $\alpha = 0.05$, I produced Δ_i for values for the difference in log likelihoods between the constrained and the unconstrained tree for each replicate. The unconstrained tree is the ML tree estimated by PAUP* from the data set. The constrained tree is constrained to the topology of T_0 before the likelihood is calculated. When performing a KH test between the tree estimated from the replicate data set (the first tree in memory) and the null topology (the second tree loaded into memory), the LnLs for trees 1 and 2 correspond to the unconstrained and constrained likelihoods. The difference in log likelihood between the two trees in the KH test is the Δ_i value that we use for the SOWH null distribution for Δ. The likelihood values are saved to file and then used to calculate the Δ_is for the Δ distribution and the log likelihood value associated with $\alpha = 0.05$ can be calculated from this distribution. For example, in the experiment SimGTRestJC there were 338 incorrectly reconstructed trees. We order the Δ_is from lowest to highest and then take the 321st (0.95x338) value as the critical value of the SOWH test. Ideally, we can work out the 95% significance level of the KH test by using 1.96x(VAR). However, a limitation of PAUP* is that the value for the variance is not an extractable output

[1] This script is also available on the accompanying CD.

value when performing a KH test. Therefore, the best we can do is find upper or lower bounds for the 95% level. The lowest significant value is the upper bound and the highest non-significant value is the lower bound. For example, the KH test for this data group showed that 1 out of the 338 trees was significantly different from the null topology. Therefore, only one Δ value is above the 95% significance level. This is the upper bound for the 95% significance level of the KH test. In the instance where zero significant results were found under the KH test, the lower bound or highest non-significant Δ is our estimate of the 95% significance level.

2.3 Results

2.3.1 The Number of Incorrectly Reconstructed Topologies Depends on Branch Lengths

The trial and error process for finding a suitable topology for generating incorrectly reconstructed topologies showed that it was difficult to produce large numbers of incorrectly reconstructed topologies for the data group simJCestJC. Leaving the topology, τ_0, and the model of evolution constant, the number of incorrectly reconstructed topologies was altered by varying the branch lengths. By introducing increasingly unequal terminal branch lengths, and by shortening the internal branch lengths of the start tree, we increase the number of incorrectly reconstructed topologies among the replicate data sets that have been parametrically simulated. It is hardly surprising that altering branch lengths has an effect, since the parametric simulation uses the branch lengths as input parameters. For example, if we take a four taxon tree ((A,B),(C,D)) with external branches of 0.1 and an internal branch of 0.1 (Figure 2.4d) and simulate sequences of 500 bp using Seq-Gen under the JC model,

we never see parametric replicates that are different from the original tree (Table 2.2). However, if we maintain finite length sequences (500bp) and the correct model, but change the tree-shape to include long branches of 0.4 and short branches of 0.01 (Figure 2.4c), the noise is sufficient to cause ML to recover the incorrect topology a certain proportion of the time (Table 2.2). The same is true for eight taxon trees (Figure 2.4a,b; Table 2.2) and clearly we can extrapolate to any number of taxa. The problem with this is that the topology has not been altered, but we have completely different results. This indicates that the power of a parametric topology test is actually dependent on more than just the topology.

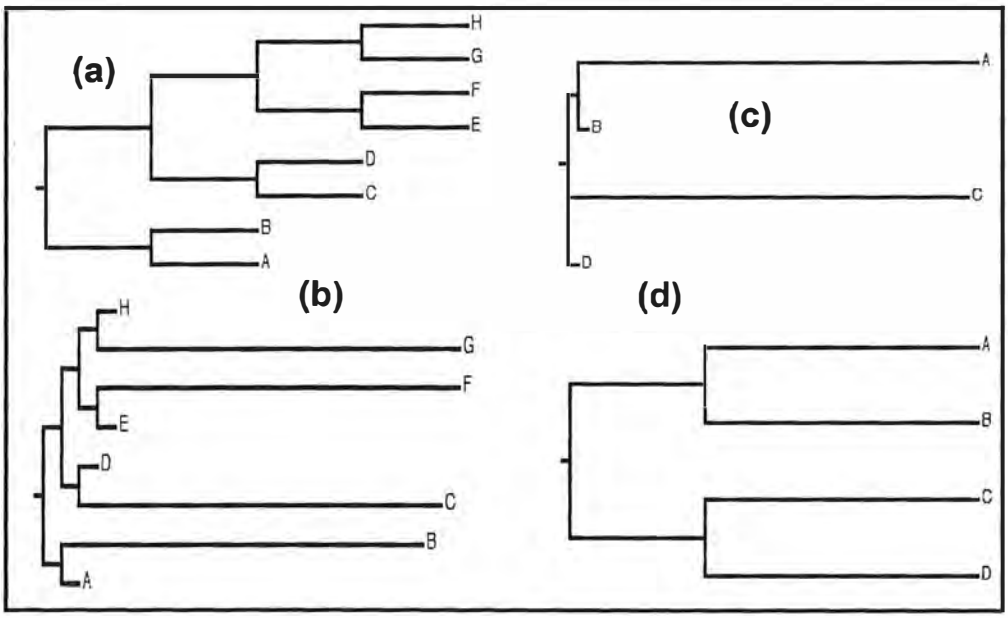

Figure 2.4 The four different trees used to illustrate that parametric simulation is not simply a test of topologies, but a test of trees.
(a) is the eight taxon tree with equal branch lengths (0.1); (b) is the eight taxon tree that has unequal branch lengths (0.02, 0.4); (c) is the four taxon tree with unequal branch lengths (0.01, 0.4); (d) is the four taxon tree with equal branch lengths (0.1).

Number of Taxa	Short Branch Length	Long Branch Length	Proportion of incorrectly reconstructed trees
4	0.1	0.1	0
4	0.01	0.4	0.175
8	0.1	0.1	0
8	0.02	0.4	0.178

Table 2.2 Summary of the effects of altering the branch lengths for a given topology.
We see that the proportion of incorrectly reconstructed trees changes from 0 to 0.175. Essentially, by increasing the ratio of the long branches to short branches on a given topology, the proportion of incorrectly reconstructed trees would tend to (1-*p*) where *p* is the probability of picking the true tree at random tree from all possible trees.

2.3.2 The Number of Incorrectly Reconstructed Topologies Depends on the Model of Evolution

The diabolical tree, T_1 (Figure 2.3), was used to simulate the data sets under six different models of evolution. The data sets from each of the six models were estimated under the same six different models of evolution. Table 2.3 shows how the number of incorrectly reconstructed topologies depends on the complexity of the models of evolution used to simulate and estimate the data.

Estimated

Simulated	JC	JC+Γ	HKY	HKY+Γ	GTR	GTR+Γ
JC	87	88	89	88	91	THTC
JC+Γ	871	416	867	402	862	THTC
HKY	201	197	118	121	117	THTC
HKY+Γ	827	587	830	463	829	THTC
GTR	338	251	280	223	141	THTC
GTR+Γ	952	639	947	585	864	THTC

Table 2.3 The number of incorrectly reconstructed topologies.
Sequences were simulated under six different models of evolution and then trees were estimated under those six different models of evolution. The rows represent the model used for simulating the sequences using Seq-Gen. The columns represent the models used to estimate the trees from those sequences. All GTR+Γ were too hard to compute (THTC) and were abandoned.

From the data presented in Table 2.3 we observe a number of effects. As the number of parameters in the rate matrix of the simulated model increase, we see that the number of incorrectly reconstructed trees increases even when the model used to

estimate the trees is the true model. For example, simJCestJC has 87 incorrectly reconstructed trees, while simGTRestGTR has (141). In addition, if the model used to estimate the trees is more complex than the true model, we observe that the accuracy of estimation is just as accurate as for the true model. This is clearest when looking at the sequences simulated under JC. All of the estimating models produce approximately the same number of incorrectly reconstructed trees. In other words, choosing a model for estimation that is more complex than the model used for simulation does not reduce the number of incorrectly reconstructed topologies (i.e. it does not increase/decrease the accuracy of estimation).

The most notable influence on the number of incorrectly reconstructed topologies is the effect of the Γ parameter. Any model for which Γ was included in simulation gave rise to significantly more incorrectly reconstructed topologies regardless of the model used to estimate the sequences. Even though including Γ is the addition of only a single parameter, estimation of the correct topology is less accurate than models that include multiple additional substitution types (e.g. HKY and GTR) even when the true model is used for estimation. To use an example, it is more difficult to accurately recover the topology for sequences simulated under JC+Γ (416) than it is to accurately recover sequences simulated under GTR (141) even when the true model is used. In addition, when estimating under models that incorporate rate heterogeneity using Γ there is no significant increase in the accuracy of estimation if the sequences were simulated under a rate homogenous model. However, the number of incorrectly reconstructed topologies for sequences simulated using a model of evolution that included Γ, is significantly reduced when Γ is included in the estimated model. This observation is similar to that of Sullivan and Swofford (2001). They showed that for

sequences that are simulated under a model that incorporates among-site rate heterogeneity, the maximum likelihood estimator is only consistent as long as among-site rate heterogeneity is accommodated in the estimating model. Since we are unable to know the true nature of evolution for any molecular sequences collected from organisms, these observations suggest large implications for incorrect model choice when estimating trees from sequences. These data suggest that estimation should always be performed with more complex models even at the expense of further computation time and that we should take into account further parameters even if these parameters are not modeled precisely (Sullivan and Swofford, 2001; Sullivan, Swofford and Naylor, 1999; Yang, 1996a; Yang, Goldman and Friday, 1994).

2.3.3 The SOWH Test versus the KH Test

To compare the SOWH and KH tests at the 95% significance level, the Δ values that correspond to the 95% significance level for each of the tests must be calculated for each data group. The Δ distributions and the two significance values associated with each test are plotted in Figure 2.5 for the data groups simGTR+ΓestJC and simHKYestJC+Γ. Figure 2.5a shows the situation where the estimated model is less complex than the simulated model. In this situation the number of incorrectly reconstructed topologies is high and the SOWH test has a 95% significance level that is higher than the KH test. Figure 2.5b shows the opposite, where the estimated model is more complex than the simulated model. Here we observe that the KH test is highly conservative with only one significant Δ value, and a higher 95% significance level than the SOWH test. Figure 2.6 shows the comparison of all the 95% significance values across the data groups for both the KH and SOWH tests. In

certain data groups, there were no significant Δ values under the KH test. In these situations the largest observed value for Δ is used to estimate the lower bound for the 95% significance value.

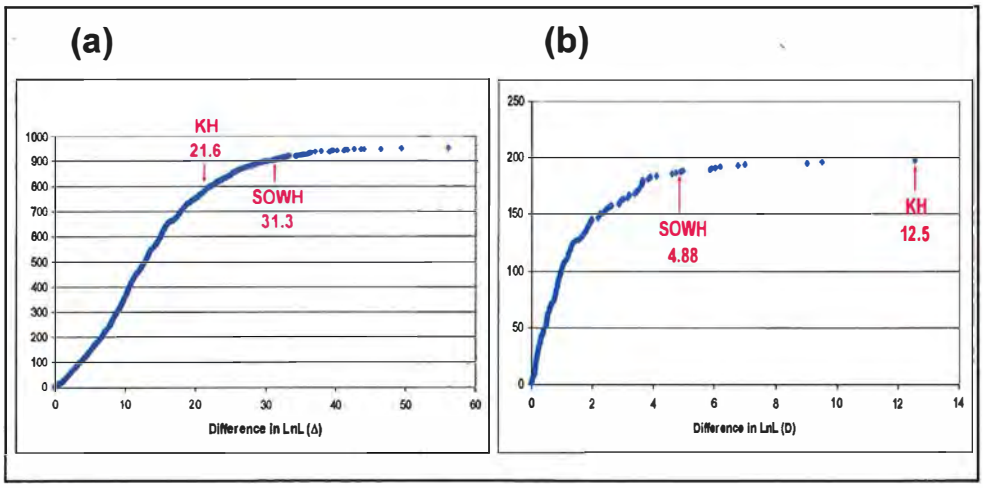

Figure 2.5 The cumulative distributions for Δ for data groups simGTR+ΓestJC and simHKYestJC+Γ.

These are cumulative distributions for the differences in log likelihood, Δ, between the constrained and the unconstrained trees for each replicate data set in the data group. The SOWH test statistic is the Δ value at $\alpha = 0.05$ (the up arrow). The 95% significance level of the KH test statistic is determined through the PAUP* results of the KH test.

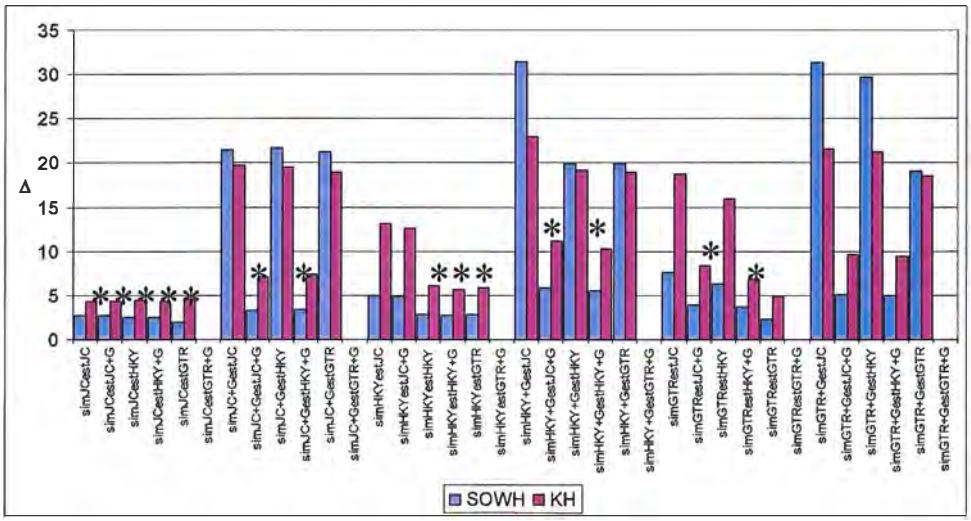

Figure 2.6 Comparison of the KH test and the SOWH test at the 95% significance level.
Under the conditions where the estimated model is insufficient to explain how the data were evolved, we observe that the SOWH test is more conservative than the KH test. However, when the model of evolution is adequately specified, we observe that the KH test is significantly more conservative than the SOWH test and in many cases unable to reject the null hypothesis at all (denoted by *).

2.4 Discussion

2.4.1 The SOWH Test is NOT Simply a Test of Topologies

Testing topologies poses many difficulties. Because the topology of a tree is discrete, it cannot be treated as a typical statistical parameter. As a result of this, it is difficult to decide whether or not there is a statistical difference between two or more topologies, because under certain circumstances different topologies may be indistinguishable from one another for the data we have gathered. Topology tests designed to give us confidence in the choice we make need to have clear and well defined hypotheses in order that they may be useful. As demonstrated by the results presented here, it is clear that the outcome of the SOWH test is influenced by both the branch lengths of the tree and the model of evolution. For tests of topology that involve the use of parametric simulation there are some serious considerations.

(1) The nature of the replicate data sets is dependent on the parameters of the assumed model, θ_0 (l_0 and M_0).

(2) The replicate data sets are actually testing the level of confidence we have in the entire tree and not just the topology.

It is well established in the literature (Antezana, 2003; Aris-Brosou, 2003; Buckley, 2002; Goldman, 1993; Goldman, Anderson and Rodrigo, 2000; Huelsenbeck and Crandall, 1997; Huelsenbeck, Hillis and Nielsen, 1996) that parametric simulation is dependent on the parameters of the assumed model, and therefore these data agree with existing data. Antezana (2003) showed specifically how the use of the parametric bootstrap generates accurate estimates of critical length only in the topological situations in which the tree used to simulate the replicates was known to

be both correct and statistically significant. This is analogous to the results in Section 2.3.1. When the topology is expected to be correct (i.e. no LBA) then we see that the correct tree is recovered. However, when there is more uncertainty in the phylogenetic estimation due to LBA, given the finite length of the sequences, we observe incorrectly reconstructed trees. To my knowledge, the SOWH test has never been referred to as anything other than a test of topologies especially when one considers that the null and alternative hypotheses for the SOWH test are: H_0: τ_0 is the true topology and H_A: some other topology is true. This, as I have shown, is not technically correct. Parametric data, by including a model, is therefore by definition not testing topologies alone. In addition, it should be noted that this doesn't make the SOWH test wrong, but its implementation should be noted as being a test of the entire tree and not simply a test of topologies.

The next problem with the SOWH test, a test of trees, has been shown to generate excessive type I error with some empirical data sets (Buckley, 2002); that is to say it is a liberal test and rejects the null hypothesis too frequently when it is true. The type I error of the SOWH test is investigated further in the next chapter, as we endeavour to understand how best to implement the parametric test to give us confidence in our tree.

2.4.2 The KH Test versus the SOWH Test

Not only is the discrete nature of topology a difficulty, but the results presented here, illustrate a further issue associated with topology testing. Two equally valid analyses, designed to test a common composite hypothesis, give quite varying results for the same data and when conditions are changed, they err in different directions. For example, when estimating T_{ML} from our biologically obtained molecular sequences

when the model is insufficient to explain the way the data were evolved, we observe that the SOWH test has a significantly higher value for Δ at the 95% significance level. This is because using an inadequate model to estimate the tree is likely to broaden what is usually a very tight distribution. In contrast, if we assume that we are estimating with a model of evolution that is a reasonable or adequate explanation of the data, the KH test is now extremely conservative, in many cases not even able to reject any of the incorrectly reconstructed trees. From the simulated data, we are actually unable to calculate the upper bound for the 95% significance level of the KH test in certain data groups. In these situations, the lower bound has been used and is represented by the largest non-significant value for Δ that we have obtained through simulation. Therefore, while we can say that the KH test is conservative under these conditions, we cannot actually quantify just how conservative it actually is.

Goldman et al. (2000) suggested that the variation in performance of the different tests may be due to the different forms of the simple null hypotheses and/or to the larger power associated with parametric tests and their stricter dependence on the substitution model. Further evidence for this is presented by Aris-Brosou (2003a; 2003b), where he shows that different tests with the same exact null hypotheses are consistent with one another in their estimation of *p*-values and therefore in their confidence set. These analyses included Bayesian analyses, bootstrap proportions, the SH test, the SOWH test and two new tests the BHT (Bayes Hypothesis Test) and the BST (Bayes Significance Test). The general conclusion is consistent with the data that we have here, in that it seems that the exact form of the null hypothesis appears to explain an important part of the difference in the results between the tests.

2.4.3 The Typical Biological Situation

The application to the typical biological situation is naturally different from a simulation study most crucially because we have no knowledge of the truth. How then should we approach the estimation of molecular sequences and the subsequent testing of topologies so that we have an accurate test? For example, if we are in a situation where the rate variation in the data is unknown, the inclusion of a rate heterogeneity parameter like Γ is imperative. Simply this is because in the event that there is rate heterogeneity, we are able to account for it. However, if there is no rate heterogeneity, the α value for the Γ distribution will be estimated at infinity. Even if α is estimated at infinity, this is still more informative about the nature of our model of evolution for those sequences than if we had not estimated α. So it seems that it would be prudent to suggest that rate variation should always be present in the estimating model. However, Yang (1997) showed a series of results in which a simpler and incorrect model that did not incorporate rate heterogeneity outperformed the true model that included rate heterogeneity. Combining these results suggests that using the model that best fits the data is the best way of estimating tree topology from the sequences, since if the model fits the data well, the SOWH test will be able to perform adequately. Selecting models of best fit for real data sets is expanded on in Chapter 4.

3 Estimation of the Type I Error Rate for the SOWH Test

3.1 Introduction

3.1.1 Overview

Estimating evolutionary histories from molecular sequences not only requires a statistical framework for estimating the trees, but also statistical tests that assess the confidence we have in the estimated tree. The elements of statistical tests that determine a choice between two competing hypotheses (e.g. the evolutionary history of our taxa) are often the error probabilities, type I and type II (Neyman, 1950). The type I error is the proportion of times the null hypothesis will be rejected when it is, in fact, true, while the type II error is the probability of accepting the null hypothesis when it is false. A good statistical test must reject the null hypothesis at the nominal value for α, which we specify in this chapter at 0.05. This is a level of significance commonly accepted as significant evidence against the null hypothesis. Tests in which the type I error exceed the pre-defined α value (e.g. type I error = 0.15) are rejecting the null hypothesis too frequently. This gives us diminished confidence in the true topology even when the null hypothesis is correct (i.e. the test is liberal). By contrast, tests that are conservative underestimate the type I error and are unable to reject the null hypothesis at the appropriate level. To accurately set the significance levels for the SOWH test we need to know the extent of the type I error. However, in a typical biological situation we have no knowledge of the true topology, the true model, or the true parameter values.

Therefore the first step when we intend to estimate type I error is to fully specify the null hypothesis (topology, branch lengths and model of evolution), so that the true relationships amongst the taxa are known. Secondly, we simulate the parametric replicate data sets that conform to the fully specified null hypothesis. Each replicate data set is now equivalent to a typical biologically obtained data set. Thirdly, each replicate data set is estimated as one would in a true biological situation, and the results are compared to the null hypothesis. Only a proportion of the replicate data sets will have incorrectly reconstructed topologies, but in every situation where the estimated tree had an incorrect topology, I performed an SOWH test. The null and alternative hypotheses of the SOWH test are:

H_0: τ_0 is the true topology

H_A: some other topology is true

For each time that we observe a significant difference between the two competing topologies, the SOWH has generated a type I error because we know that the null hypothesis is true. By counting the number of times that we observe a type I error in each experimental data group, two probabilities associated with the SOWH test can be calculated. Firstly, we calculate the marginal probability as the proportion of times that the SOWH test rejects the null hypothesis out of the total number of replicates that were generated. Secondly, we calculate the conditional probability as the proportion of times that the null hypothesis is rejected out of the number of datasets that estimated incorrect topologies. Both probabilities are informative in different ways and how they change with respect to one another is of interest to us.

3.1.2 δ, δ` and the Δ distribution

In practice, the SOWH test, like other tests of topology, is only performed when we obtain an unexpected result. A typical situation arises when we obtain a topology that is different from the well established topology after performing a maximum likelihood estimation on our data. In other words the topology of T_{ML}` is not the same as τ_0. It is in performing simulation studies of the SOWH test that the three most fundamental terms and symbols in this chapter arise. These are the three different 'deltas'. As with previous uses of delta so far in this thesis, it is the difference between two log likelihoods from different topologies for the same data set. In this chapter that is no different, but the subtle differences require some elucidation. Little delta, δ, and little delta prime, δ`, are the test statistics of an SOWH test. The difference between each of these is the number of constraints placed on them. δ is the value for the difference in log likelihoods between τ_0 and τ_{ML} when the parameters of the model, are not estimated, but are equal to the parameters of the null hypothesis. This is referred to as the true value for δ because it uses the exact specifications of the null hypothesis. δ` is the estimate of δ and is the difference in log likelihoods between τ_0 and τ_{ML} when only the topological constraints are enforced and the parameters are estimated. This is equivalent to a delta value that is obtained from real data where one has no knowledge of the true value for δ.

The first part of this chapter compares δ to δ` to illustrate the accuracy of the δ` values, because erroneous estimates of δ may compound to cause unpredictable behaviour in the latter stages of the analysis. Therefore our questions of interest are:

(1) How good an estimate is δ` for δ?

(2) What tree shapes or model assumptions affect accurate estimates of δ?

(3) How large or small are these inaccuracies?

The second part of this chapter estimates the type I error of the SOWH test. For each data set where $\tau_{ML} \neq \tau_0$ the SOWH test is performed. 1000 replicate data sets were simulated and trees were estimated for each of these data sets as per a normal SOWH test. For each data set, the difference between LnL_{ML} and the log likelihood of the data constrained to τ_0, Δ_i, is calculated. Collectively the Δ_i's make up the Δ distribution against which we test our test statistic, δ`. Two probabilities associated with the test (the marginal and constrained probabilities for rejecting the null hypothesis) can then be calculated using the proportion of times that the SOWH test rejects the null hypothesis, because we know that the null hypothesis is in fact correct. We have predefined our α value for the SOWH test at 0.05, a value that most researchers and publishers accept as "strong evidence" against the null hypothesis. Therefore, if δ` is greater than 95% of the values in the Δ distribution then we have rejected the null hypothesis.

3.2 Methods Part 1

3.2.1 Defining the Null Hypothesis Trees

I have defined five null hypothesis trees as the true trees for this chapter (Table 3.1). For this chapter four taxon trees have been used because of the significant increase in computation that is anticipated in the two rounds of estimation. The first round is the estimation of the ML tree from the original 1000 replicate data sets and the second is then performing the SOWH test on each one of those data sets where the ML tree is

not the true tree. Figure 3.1 shows the four taxon topologies of the null hypothesis trees that were used for simulation. In all trees except the star tree (zero internal branch length), the internal branch was fixed at 0.01. The 'Felsenstein' tree (tree 'a') has two long branches and two short branches. The long branches are set to 0.3 on two non-adjacent taxa and the short branches are set to 0.01 (also on non-adjacent taxa). This tree-shape was chosen because of the known long-branch attraction (LBA) that occurs during estimation for finite data sets. The LBA causes taxon A and taxon B (the unrelated taxa on the long branches) to appear more closely related leading to the reconstruction of an incorrect topology. The 'Farris' tree (tree 'b') also has two long branches (0.3) and two short branches (0.01). However, they are partitioned differently by the internal branch leaving one long branch most closely related to one short branch. It was claimed that parsimony 'out-performs' ML in the case of the Farris tree (Siddall, 1998) with ML being subject to "long-branch repulsion[1]", but this is only due to the fact that parsimony is not a consistent estimator of topology in some cases (Swofford et al., 2001) and in this situation it is actually biased towards being correct more often than it should be.

Name	**Terminal *l***	**Internal *l***	**Newick Tree**
Felsenstein	0.3 and 0.01	0.01	((A:0.01,B: 0.3):0.01,(C:0.3,D:0.01):0)
Farris	0.3 and 0.01	0.01	((A:0.01,B:0.01):0.01,(C:0.3,D:0.3):0)
Equal0.1	0.1	0.01	((A:0.1,B:0.1):0.01,(C:0.1,D:0.1):0)
Equal0.3	0.3	0.01	((A:0.3,B:0.3):0.01,(C:0.3,D:0.3):0)
Star	0.3	0	((A:0.3,B:0.3):0.00,(C:0.3,D:0.3):0)

Table 3.1 The five trees used to simulate the sequences in chapter 3.

[1] Long-branch repulsion is not a term that is advocated by me nor work performed in this thesis. However, for consistency through the literature this term is used occasionally when considering the noted behaviour of ML estimation of data simulated using a Farris tree.

A tree with equal terminal branch lengths, shown in Figure 3.1c, was used for two trees. The terminal branch lengths for one of the trees were 0.3 and the other had terminal branch lengths of 0.1. The fifth tree also had all branch lengths equal to 0.3, but had no internal branch. This star topology is a special class of all three possible topologies for the four taxa (Antezana, 2003) and was included for completeness.

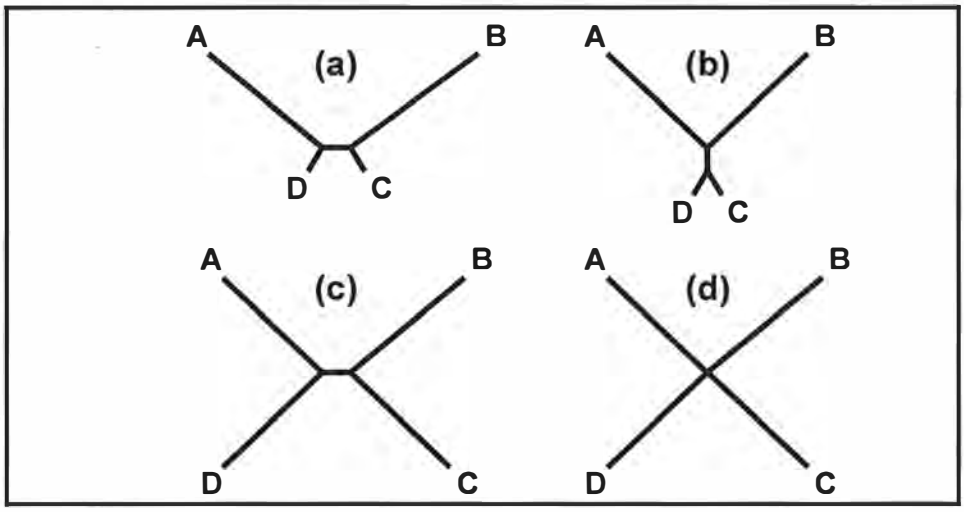

Figure 3.1 The null hypothesis trees used for simulation in chapter 3.
The three bifurcating topologies and the star topology are shown for four taxa. (a) is termed a Felsenstein tree; its dominant feature is LBA between taxon A and taxon B. (b) is termed the Farris tree; its dominant feature is the apparent "long-branch repulsion" that exists under likelihood. (c) is a tree with equal terminal branch lengths. (d) is the star topology with a zero internal branch length and equal terminal branch lengths.

3.2.2 Simulating and Estimating the Replicate Data Sets

1000 replicate data sets were simulated according to six different models of evolution (JC, JC+Γ, HKY, HKY+Γ, GTR and GTR+Γ) using Seq-Gen for each of the null hypothesis trees (Figure 3.2). In the more complex models, the parameters for nucleotide frequencies (HKY and GTR) and the rate matrix (GTR only) are empirical estimates for the HIV-1 *env* gene (Table 3.2). For each tree under each model, the length of nucleotide sequence was varied to investigate the impact that this may have on parametric simulation. There were three categories of nucleotide length 250, 500 and 1000 bp.

Models	Seq-Gen Command
JC	-l 500 -on -n 1000
JC+Γ	-l 500 -on -n 1000 -a 0.4
HKY	-l 500 -on -n 1000 -m HKY -t 2 -f0.4 0.15 0.21 0.24
HKY+Γ	-l 500 -on -n 1000 -m HKY -t 2 -f0.4 0.15 0.21 0.24 -a 0.4
GTR	-l 500 -on -n 1000 -m REV -r 0.018 0.17 0.52 1.22 3.08 1 -f 0.4 0.15 0.21 0.24
GTR+Γ	-l 500 -on -n 1000 -m REV -r 0.018 0.17 0.52 1.22 3.08 1 -f 0.4 0.15 0.21 0.24 -a 0.4

Table 3.2 The models used to simulate the replicate sequences using Seq-Gen.
-l sets the length of the sequence (either 250, 500 or 1000). -on sets the output file to be in nexus format. -n specifies the number of replicate data sets to simulate. -m specifies the model of evolution to use (JC [default], HKY or REV). -t sets the value of the ratio of transitions to transversions. -f sets the values of the initial base frequencies. -r sets the parameters of the rate matrix (a-f). -a sets the alpha value of the gamma distribution.

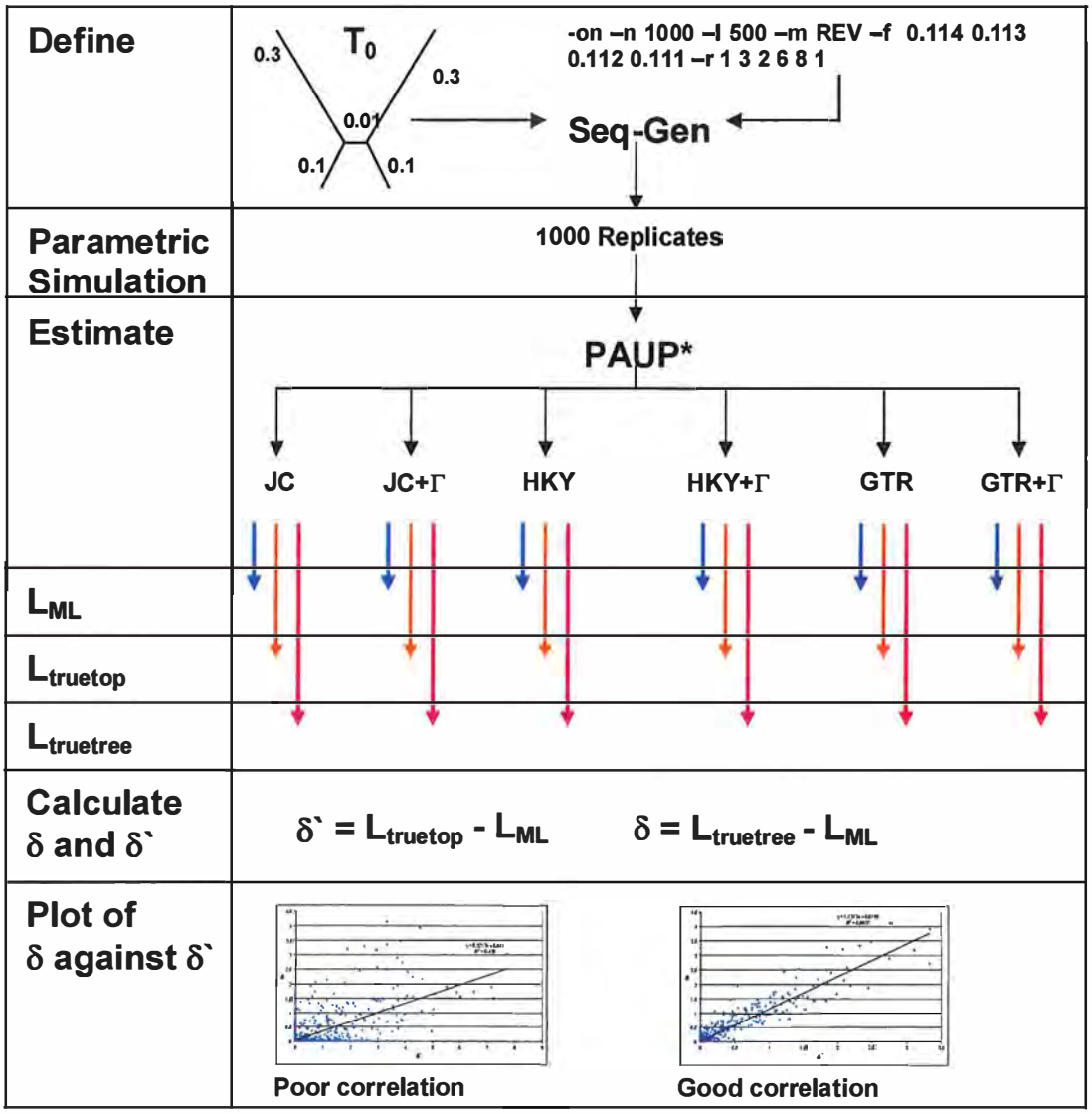

Figure 3.2 A diagram of the procedure for generating the plots of δ` against δ.
The blue arrows represent unconstrained maximum likelihood estimation, the orange arrows represent the estimation of the likelihood for the data constrained to the null topology, τ_0, and the red arrows represent the estimation of the likelihood for the data constrained to the null topology and the parameters of the model, θ_0.

Each simulated data set was estimated under each of the six different models of evolution to form a 'data group'. The experimental notation for a data group that consists of 1000 data sets is 'Tree simMODEL$_1$estMODEL$_2$ seq-length bp' (e.g. Farris simHKYestJC+Γ 500 bp). Each data group was analysed by three different estimation procedures.

(1) The unconstrained maximum likelihood was estimated from the data to give T_{ML} and therefore L_{ML}.

(2) The likelihood was estimated for the data constrained to the null hypothesis topology and the parameters were estimated for each of the six different models to give $L_{truetop}$[1].

(3) The likelihood was estimated for the data constrained to the null tree (including the branch lengths, l_0) and the null hypothesis parameters of the true model of evolution, M_0 to give $L_{truetree}$[2].

This required a combination of Perl scripts[3]. First, a single file of PAUP commands was generated in which the entire estimation process was automated using CONCAT4PAUP. Then, the log files were divided into their experimental data groups with DIVIDOR4EXTRACTOR. Finally, the δ and δ' values were calculated

---

[1] $L_{truetop}$ is the likelihood of the data when it is constrained to the true topology

[2] $L_{truetree}$ is the likelihood of the data when it is constrained to the true tree and the true model

[3] The scripts CONCAT4PAUP, DIVIDOR4EXTRACTOR, ULTIMATE EXTRACTOR and tractorSUB-X were used in this section and are available on the accompanying CD.

and extracted in different ways using ULTIMATE EXTRACTOR and tractorSUB-X. to be imported into Microsoft Excel for further analysis.

3.2.3 Comparing δ` Estimates to the True Value of δ

540 different data groups were created that cover all combinations of 'Tree simMODEL$_1$estMODEL$_2$ seq-length bp', where 'Tree' is one of the five null hypothesis trees, 'simMODEL$_1$' is one of the six models of evolution used to simulate the data, 'estMODEL$_2$' is one of the six models of evolution used to estimate the data and 'seq-length' is one of the three different sequence lengths of the simulated data. The LnLs that correspond to L_{ML}, $L_{truetop}$ and $L_{truetree}$ have already been estimated and extracted appropriately so that LRTs that correspond to δ and δ` could be constructed. In any instance where the estimated maximum likelihood topology is different from the true topology, $L_{ML}$ will have a different value to $L_{truetop}$. δ`, the difference between L_{ML} and $L_{truetop}$, was calculated for each data set. The true value for δ is the difference between L_{ML} and $L_{truetree}$. δ` was calculated for every replicate data set estimated under the six different models of evolution. If δ` is an accurate estimate of the true value of δ, then we expect that on average δ` = δ. I constructed *x,y* scatter plots comparing δ to δ` to illustrate the fit of δ to δ`. If they are approximately equal this will produce a straight line with a slope approximately equal to 1 with a high Pearson's correlation coefficient ($r^2$). A poor correlation (low $r^2$) between δ and δ` indicates that δ` is not a good estimate of δ. Erroneous values of δ`, merely through assumptions that we make at the beginning of an analysis when estimating a tree from the sequences may lead to incorrect testing under the SOWH test. 540 scatter plots (one for each data group) comparing every δ` to its corresponding δ value were created and analysed.

3.3 Results Part 1

3.3.1 The Parameters that Affect the Accuracy of δ`

Each of the 540 x,y scatterplots of δ` against δ were visually inspected. Lines of best fit and the correlation coefficients ($r^2$) for the data were calculated and displayed on the graphs. We are interested in the values for the slope and the $r^2$. A slope on the graph that is approximately equal to 1 indicates that δ` and δ are approximately equal, while the r^2 value measures how highly correlated the data are. A high r^2 (close to 1) indicates that the data correlate well to one another. From these analyses, it is clear that δ` is robust to certain aspects of the null hypotheses, while greatly affected by other parameters. Across all the experiments, sequence length did not affect the fit of δ to δ` an example of which is shown in Figure 3.3.

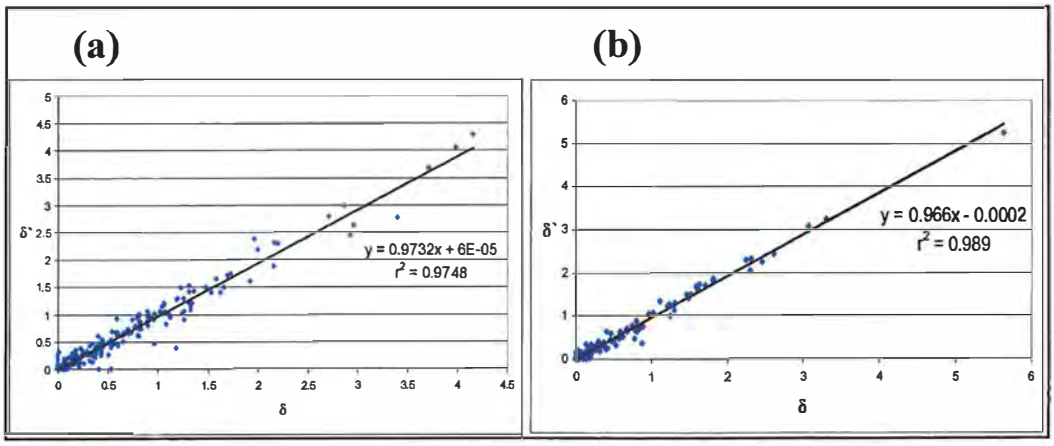

Figure 3.3 x,y scatterplots of δ`against δ for Felsenstein simHKYestHKY with different sequence lengths.
The data sets used in experiment (a) were 250bp long while those in (b) were 500bp long. The trends on the graphs are almost identical and illustrate that sequence length has no effect on the capacity of δ` to estimate δ`.

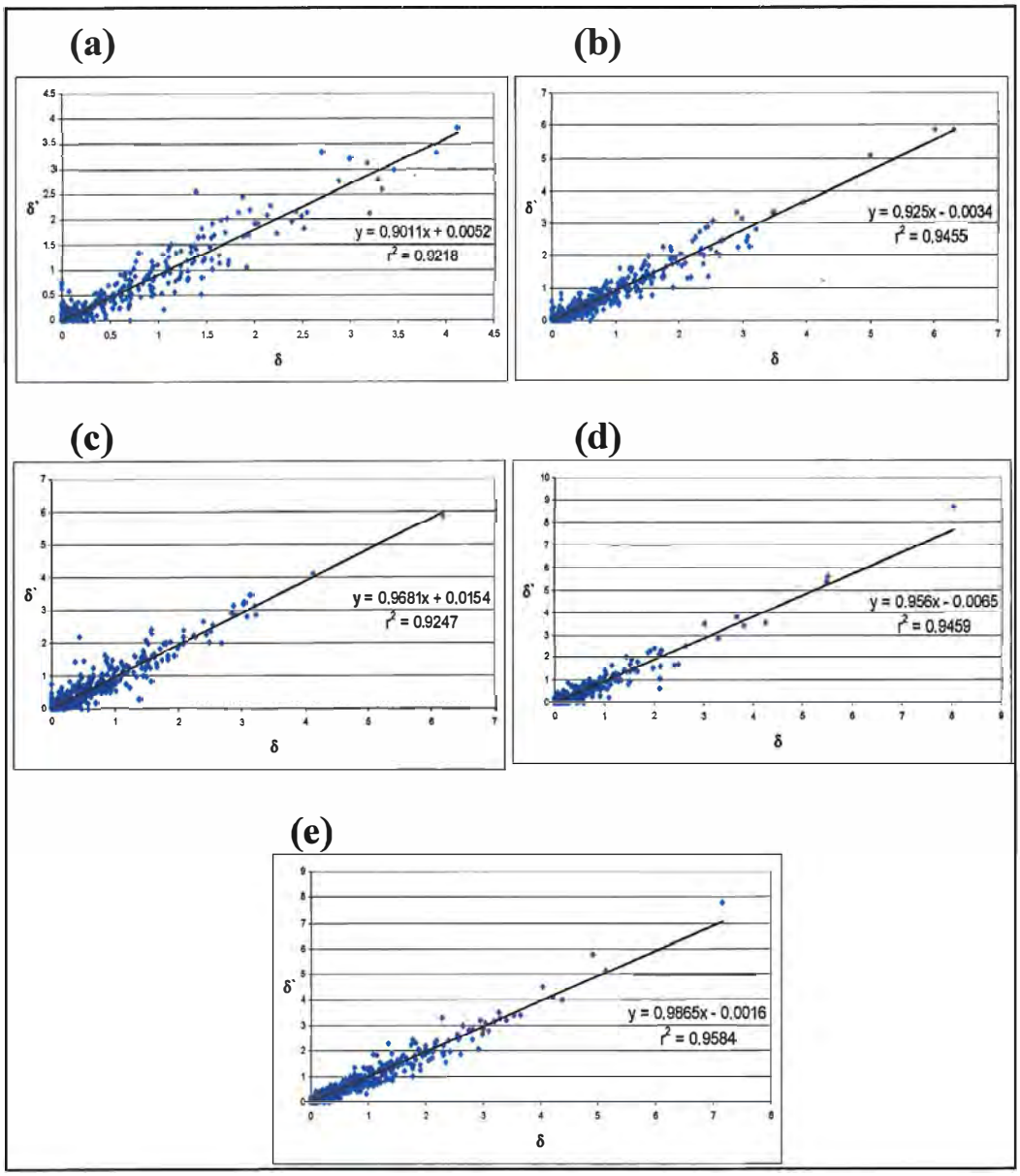

Figure 3.4 x,y scatterplots of δ` against δ for sufficiently complex models on all trees.
(a) Experiment Equal0.1 simGTRestGTR 250 bp. (b) Experiment Equal0.3 simGTRestGTR 250 bp (c) Experiment Farris simGTRestGTR 250 bp. (d) Experiment Felsenstein simGTRestGTR 250 bp. (e) Experiment Star simGTRestGTR 250 bp. For each of these the slope is approximately 1 and the r^2 is high. This collection of graphs also illustrates that tree type does not affect the relationship between δ and δ`. The similarity between all the graphs illustrates that the relationship between δ and δ` is independent of tree type.

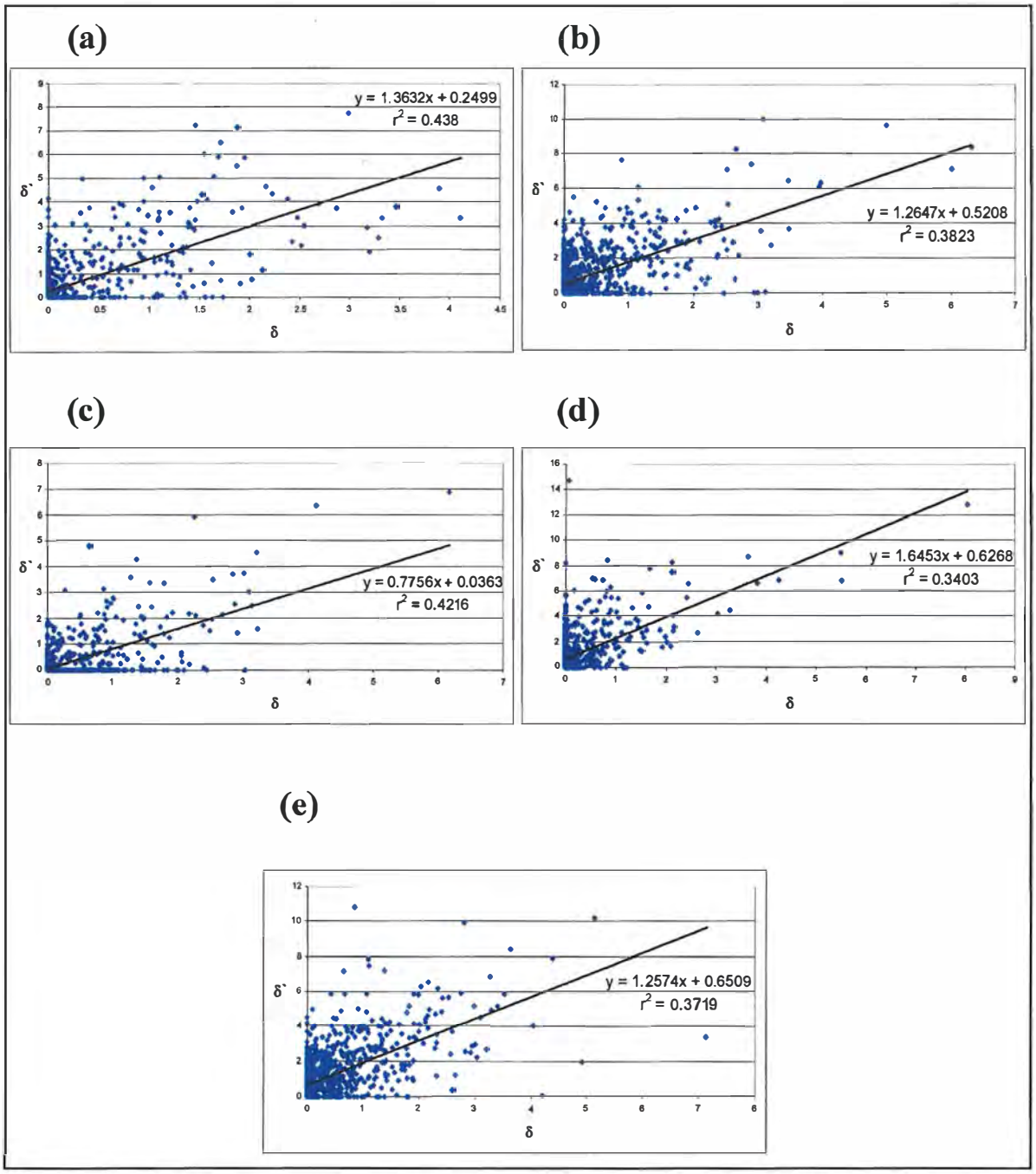

Figure 3.5 *x*,*y* scatterplots of δ` against δ for insufficiently complex models on all trees.
(a) Experiment Equal0.1 simGTRestJC 250 bp. (b) Experiment Equal0.3 simGTRestJC 250 bp (c) Experiment Farris simGTRestJC 250 bp. (d) Experiment Felsenstein simGTRestJC 250 bp. (e) Experiment Star simGTRestJC 250 bp. For each of these data groups the slope indicates that δ` is an overestimate of δ. However, the low $r^2$ value shows that this is not a particularly strong correlation. The similarity between all the graphs once again informs us that the relationship between δ and δ` is independent of tree type.

The major influence on δ across all experiments is the complexity of the estimated model relative to the simulated model. Figure 3.4 shows that δ` is a good estimate of δ across all five tree shapes when the estimated model is sufficiently complex to explain the model under which the data were simulated. The results in Figure 3.4 can be contrasted with the results in Figure 3.5 to show that when the model used to estimate the data is not sufficiently complex, δ` is not a good estimate of δ. At the same time, because the trend is consistent across all tree types, these results suggest that the tree type used to simulate the data also has no effect on the capacity of δ` to estimate δ. In Figure 3.5, all the data groups have a trendline with a gradient of less than one. Therefore, we can say that δ` appears to be an overestimate for δ. However, we should reserve a sweeping extrapolation from the data presented because the r^2 values show that the fit to the data is not particularly strong.

3.4 Discussion Part 1

3.4.1 Model Effects in Estimation of δ`

These results show that the dominant causes for inaccurate estimates of δ were all associated with the estimated model of evolution and not the length of the simulated sequences nor the tree type used to simulate the sequences. This has implications for the choice of the model used for the estimation of trees from sequences. When the model of evolution used to estimate the tree from the sequences is not sufficiently complex, we observe a weak trend that shows δ` generally has a larger value than δ as shown by the value for the slope and therefore δ` is not an accurate estimate of δ. However, if the model of evolution used to estimate the sequences is sufficiently complex, we observe that the slope is approximately equal to one (i.e. $\delta` \approx \delta$) and that

the r^2 is high which indicates this correlation is significant. Therefore, under these circumstances δ` is a good estimate of δ. Inflated estimates of δ through the choice of an insufficiently complex model when estimating the tree, may contribute to the higher number of observed rejections of the null hypothesis when it is true (i.e. increased type I error). Huelsenbeck, Hillis and Nielsen (1996) showed a similar result in their LRT of monophyly where parametric bootstrapping was robust to changes in topology, but that constricting assumptions in the substitution model caused the method to fail.

Calculating δ is simply the first part to the grander scheme of performing an SOWH test on a topology that is different from τ_0. The data presented in Figures 3.4 and 3.5 suggest that in the event that we require an SOWH test, our estimate of δ will be more accurate if the initial estimation of the tree was performed using a more complex model of evolution. Therefore, a general suggestion to always use a more complex model seems reasonable.

3.5 Methods Part 2

3.5.1 Performing the SOWH Test on the Replicate Data Sets

The SOWH test and the estimation of its type I error was not performed on all the data groups for which δ` values had previously been calculated. The computational intensity allowed for performing these tests on only 18 out of the 540 data groups. These are the data groups of 500 bp simulated using Farris and Felsenstein trees under JC, HKY and GTR. Figure 3.6 shows the entire procedure for assessing type I error. The first four steps (i.e. up to calculating values for δ`) have already been performed as part of generating a comparison of $\delta$` and δ and are the same steps as

shown in Figure 3.2 and as described in Section 3.2.2. 1000 parametric replicates were simulated using Seq-Gen under the estimated model of evolution and the null hypothesis tree as the start tree. As per a normal SOWH test, trees were estimated for each of the replicates and Δ_i values were calculated for the difference between the likelihoods of the constrained and unconstrained trees. This procedure is analogous to the iterated or double bootstrap (Efron, Halloran and Holmes, 1996; Hall and Martin, 1988; Rodrigo, 1993)

3.5.2 The Analysis of Type I Error for the SOWH Test

Type I error is the probability of rejecting the null hypothesis when it is true. Here we know the null hypothesis, and a certain number of the data sets within our data groups estimated the incorrect topology. Our null hypothesis for the SOWH test is that the true topology, τ_0, is shown statistically to be true. If it is not, then we are making a type I error by rejecting it. The 95% significance level of the SOWH test is set by ordering the Δ_i values and taking the 950^{th} (0.95 x 1000) as the cut-off value. We can then count the number of times that δ` is rejected to get an estimate of the type I error. Alternatively, we can compare $\delta$` to each Δ_i value and count the number of times that δ` is greater than Δ. We count the number of times that this number is greater than fifty (0.05x1000) to get an estimate of the type I error. This procedure is implemented in a Perl script (d2alpha.pl[1]).

[1] d2alpha.pl is available in the appendix and on the accompanying CD

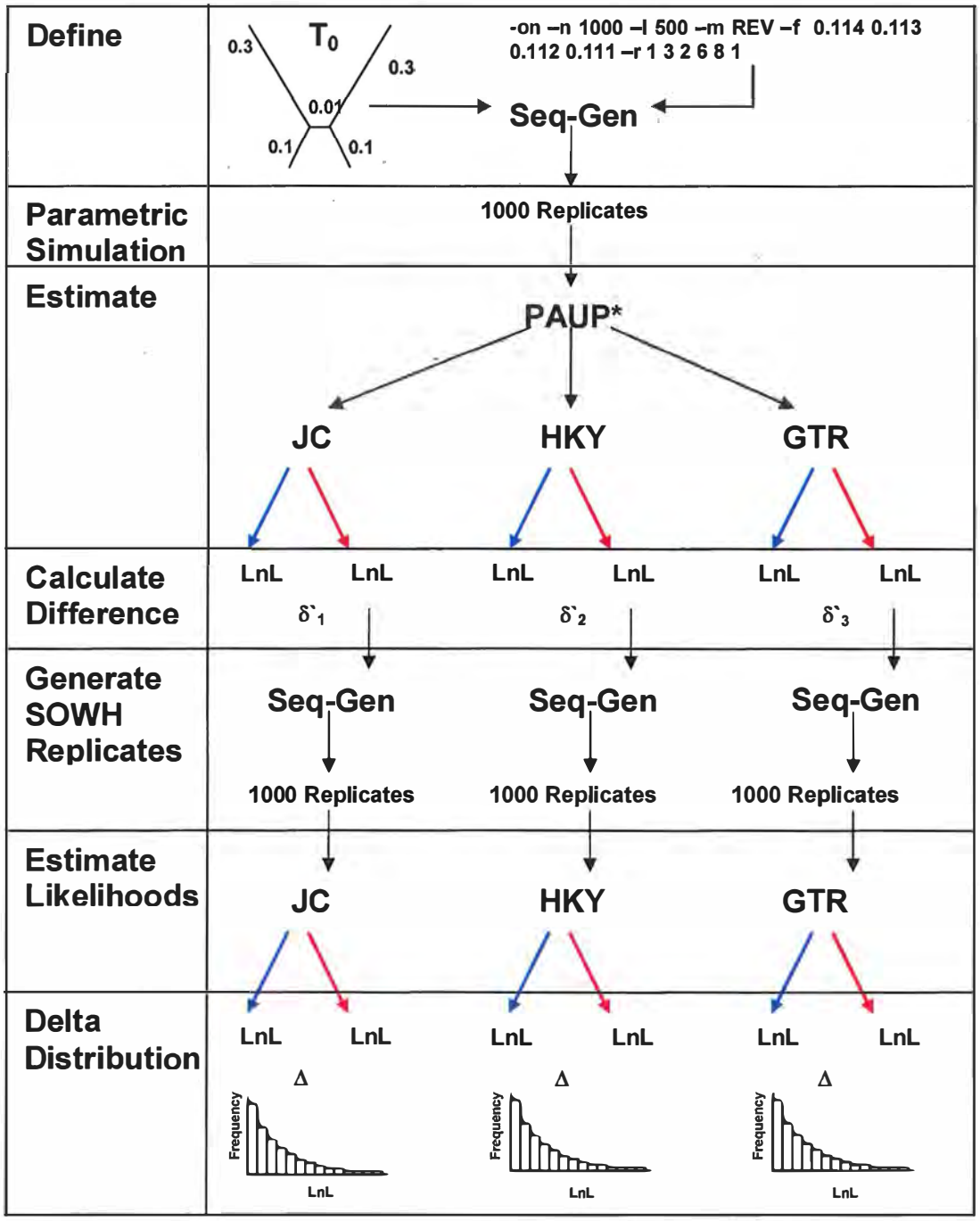

Figure 3.6 The complete flow diagram for assessing the type I error of the SOWH test. This was performed using only the Farris and Felsenstein trees as the null hypothesis trees. The sequence length was 500 bp and the simulated models defined in the top right corner correspond to the Seq-Gen commands in Table 3.2 for the JC, HKY and GTR models. The estimated models are those shown here (JC, HKY and GTR). Red arrows indicate that the estimation was constrained to the null hypothesis topology, while blue arrows are an unconstrained ML estimation. δ` values were tested against the Δ distribution to estimate the type I error.

3.6 Results Part 2

3.6.1 The Probabilities Associated with the SOWH Test

The type I error is the proportion of times we reject the null hypothesis when it is true. The null hypothesis is known to be true and therefore any instance that we observe a rejection of our null hypothesis, we are making a type I error. There are two probabilities that we are able to calculate for the 18 data groups that were tested using the SOWH test. The marginal probability for the type I error, is the total number of significant results divided by the number of replicate data sets, which in these simulations was always 1000 (Eq. 3.1). The second probability is a conditional probability that is conditioned on the fact that the tree estimated from that data set must have a topology different from that of the null hypothesis (Eq. 3.2).

$$P(marginal) = \frac{\#\,significant}{\#\,replicates} \qquad \text{Eq. 3.1}$$

where #significant is the number of significantly different results according to the SOWH test and #replicates is the number of replicates in the data group (always 1000).

$$P(conditioned) = \frac{\#\,significant}{\#\,incorrect} \qquad \text{Eq. 3.2}$$

where #significant is the number of significant different results according to the SOWH test and #incorrect is the number of incorrectly reconstructed topologies amongst the 1000 replicate data sets in the data group.

The number of incorrectly reconstructed topologies and the number of significant results for data simulated using both Felsenstein and Farris trees for 500 nucleotides

were counted by my Perl script (d2alpha.pl) and are displayed in Table 3.3 (Felsenstein tree) and Table 3.4 (Farris tree). Both the marginal and conditioned probabilities for the type I error associated with the SOWH test are also presented in these tables.

Model	#significant	#incorrect	P(conditioned)	P(marginal)
simGTRestGTR	9	121	0.074	0.009
simGTRestHKY	271	421	0.644	0.271
simGTRestJC	353	491	0.712	0.353
simHKYestGTR	18	126	0.143	0.018
simHKYestHKY	17	128	0.133	0.017
simHKYestJC	144	292	0.493	0.144
simJCestGTR	35	105	0.333	0.035
simJCestHKY	18	104	0.173	0.018
simJCestJC	14	113	0.124	0.014

Table 3.3 The type I error for the 9 data groups simulated with a Felsenstein tree.
#incorrect is the number of incorrectly reconstructed topologies during the estimation. The SOWH test was performed on each of these, and the number that were significant under the SOWH test are recorded in the #significant column. The proportion of #significant / #incorrect is the type I error.

The results of the Felsenstein tree experiments in Table 3.3 illustrate a few phenomena that contribute to the observed type I error values. On the whole, these results are in concordance with the general point made by Shibata (1989) that model underfitting is potentially more of a problem than overfitting. We observe that the number of incorrectly reconstructed topologies does not change significantly if the model used to estimate the sequences is at least as complex as the model used to simulate those sequences. For example, all three models of evolution used to estimate the sequences simulated under JC have approximately the same number of incorrectly reconstructed trees. This is because all the models are sufficiently complex to explain how the sequences were simulated for approximately 90% of the data sets. This holds true for the sequences simulated under the HKY model of evolution. We see that the numbers of incorrectly reconstructed trees are similar when the data are estimated

under the HKY and GTR models. However, when the sequences are estimated under JC, we see that the number of incorrectly reconstructed topologies increases approximately 2.5-fold. Therefore, it appears that underfitting the model is far worse than overfitting the model. For the sequences simulated under GTR, we observe that both data groups, where the estimated models were underfitted for the data (HKY and JC), there is approximately a 3.5-fold increase in the number of incorrectly reconstructed trees than those estimated under GTR.

Model	#significant	#incorrect	P(conditioned)	P(marginal)
simGTRestGTR	42	405	0.104	0.042
simGTRestHKY	57	134	0.425	0.057
simGTRestJC	49	106	0.462	0.049
simHKYestGTR	47	408	0.115	0.047
simHKYestHKY	46	396	0.116	0.046
simHKYestJC	69	195	0.354	0.069
simJCestGTR	47	377	0.125	0.047
simJCestHKY	19	359	0.053	0.019
simJCestJC	50	414	0.121	0.05

Table 3.4 The type I error for the 9 data groups simulated with a Farris tree.
#incorrect is the number of incorrectly reconstructed trees during the estimation. The SOWH was performed on each of these, and the number that were significant under the SOWH test are recorded in the #significant column. The proportion of #significant / #incorrect is the type I error.

However, of most significant interest is the type I error probabilities (both marginal and conditioned) and the reasons for their increase. The three most significantly inflated type I error values (simGTRestHKY, simGTRestJC and simHKYestJC) are those data groups in which the model is insufficient to estimate an accurate value for δ. The δ` values in these data groups are larger than true values for $\delta$, as shown in Section 3.3.1 and as discussed in Section 3.4.1. The increased value for $\delta$` is quite likely to account for the increased number of rejections of the null hypothesis, as it is more likely to be greater than 95% of the values in the Δ distribution. Oddly, we see that in data group simJCestGTR that approximately twice as many data sets rejected

the null hypothesis than for the other two simulated models. This spurious result has been double-checked and remains unexplained.

The data in Table 3.4 illustrates marginal probabilities that are approximately the same in 8 out of 9 of the data groups for the Farris tree. Interestingly, the marginal probabilities are very close to the nominal α value of 0.05. Yet, the conditioned probabilities for the SOWH test vary markedly between the data groups where the estimate model was at least as complex as the simulated model and the data groups where the models were less complex. It is known that in the anti-Felsenstein zone (two long adjacent branches on a four taxon tree - also termed Farris zone and inverse-Felsenstein zone) simpler models are biased to be more correct during estimation (Siddall, 1998; Swofford et al., 2001). Therefore, it is not too surprising that the number of incorrectly reconstructed trees is greater when the estimated model is as complex as the simulated model. Although we would expect that inflated δ` values should also play a role in the number of rejections of the null hypothesis, this does not seem to be the case for the Farris tree.

3.7 Discussion 2

3.7.1 The Marginal and Conditional Probabilities

The two probabilities associated with the SOWH test reflect two different ways of looking at the situation of hypothesis testing of topologies. In a classical sense, the marginal probability represents the scenario where any tree that is ever estimated is included for testing of the topologies. However, due to the discrete nature of topologies, we can readily identify when two topologies are identical and therefore do not strictly require testing, because the result is by definition already known. Two

identical topologies cannot possibly be found to have a statistically significant difference, because they are not different. While this assessment is biologically acceptable, with respect to the results presented here, the conditional testing of only topologies that are different from one another causes the inflation of the type I error for the SOWH test.

If one looks at a typical biological situation where topologies are only tested when the topology of the estimated tree is different from that of the null topology, we have instantly conditioned our test. Therefore, the true biological type I error is the conditioned probability and its behaviour is of concern to us. While the proportion of significant results remains roughly the same (as determined through the marginal probability) we observe that the conditional probability changes significantly depending on the nature of the estimated and simulated model. The results of the Farris tree experiment illustrate that the marginal probability is approximately the same in eight of the nine data groups (the model does not seem to have a significant effect). However, when the number of significant results is conditioned on the number of incorrectly reconstructed topologies, the effect of model complexity is profound. In the situation where the model is sufficiently complex to explain how the sequences were evolved, the number of incorrectly reconstructed topologies is larger than if the model were simpler and incorrect. Once again we are forced to consider the benefits of selecting a model of best fit for the data, given that model selection is again of the utmost importance and that the choice of model has a profound effect on the type I error.

These results agree with the simulation study of Huelsenbeck, Hillis and Nielsen (1996) where they showed that their parametric LRT of monophyly should be

performed with models of DNA substitution as parameter rich as possible, or else one runs the risk of a high level of type I error. Therefore, it seems that the type I error under model misspecification is at an unacceptable level for the SOWH test when the model is underfitted. However, if we have made every effort when estimating the trees from the sequences and when performing the SOWH test to use models that encompass the complex features of the biological data, we anticipate that results obtained using the SOWH test are a valuable step for any phylogenetic analysis.

4 Site-Pattern Analysis of Observed Data

4.1 Introduction

4.1.1 Outline

In this chapter, the site-patterns of five data sets are analysed. The five data sets are aligned molecular sequences from HIV-1, Primates, Rodents, Bees and Birds. Site-patterns are the fundamental unit of aligned sequences from which phylogenetic trees are built under maximum likelihood. Artefacts in the molecular sequences may cause the models to incorrectly recover the site-patterns under parametric simulation and lead to incorrect testing of phylogenetic data. For each of the HIV-1, Primate, Rodent and Bird data sets there is a topology that is known to be well-supported by numerous data sources. Of those data sets presented here, the Rodent and Primate data sets each estimate the well-supported hypotheses under sufficiently complex models of evolution, but the HIV-1 and Bird data sets do not. There are biases in the HIV-1 and Bird data sets that are known to cause violations in the underlying assumptions of estimating trees from sequences (e.g. i.i.d. or the use of a single model of evolution to explain the evolutionary history of concatenated sequences). The Bee data set does not have a known well-supported topology, and is an example of a data set that contains conflicts in the data that may be better explained by a network rather than a strictly bifurcating tree. These data sets provide us with a range of empirical examples with which we can assess the ability of the parametric bootstrap to recover site-patterns. It is more important to verify if models can accurately recover the data than if they can recover the estimated tree from the replicate data sets because assumptions

about the data made during the tree building process have not yet been. Using the model of best fit, we can parametrically simulate data sets that are samples from the distribution of that model. If the observed data is atypical of the distribution expected under the given model and topology, there are flaws associated with the model. Various hypotheses are tested and possible explanations are proposed for each of the data sets with respect to the results of the site-pattern analyses.

To perform the entire procedure on each data set we require:

(1) A program to estimate likelihood scores for 56 nested models of evolution on a Neighbour-Joining tree (PAUP*)

(2) A program that chooses the single model that best fits the data based on the likelihood scores (Modeltest).

(3) A program to estimate the parameters for the model of best fit (PAUP*).

(4) A program to simulate the parametric replicate sequences (Seq-Gen).

(5) A Perl script that counts the site-patterns in the observed data and in each of the replicates.

(6) Another Perl script that automatically generates the values for the null distribution of expected site-pattern variation.

(7) A test of the value obtained from the observed data against the expected distribution to determine if it is typical or atypical of the null distribution.

The result obtained from testing the observed sequences against the expected distribution gives us an indication of how typical our observed data is. Artefacts that

favour non-treelike structures (recombination, reassortment and horizontal transfer) and model misspecification may manifest in the way parametric simulation recovers the site-patterns. Therefore, it is important to know whether or not the site-patterns that are present in the observed data are recovered correctly in parametric simulation. In this chapter it was shown that the ability of parametric simulation to recover site-patterns was data dependent. Data sets that appear well explained by a single model of evolution and estimate the "true" tree, have less variation in their site-patterns than 95% of the values in the expected distribution. However, where the model is badly violated, as in the Bird data set, we observe an atypical result. This suggests that the site-pattern recovery is poor and possible explanations are provided in the discussion. Alternative ways of breaking down the data for more accurate and biologically relevant analyses are discussed.

4.1.2 Non-treelike Influences and Heterogenous DNA

Data simulated under a tree, T_0, is done under the assumption that the information in the molecular sequences is "treelike". However, treelike structures are not necessarily always the best way of representing evolution. For example, where recombination, reassortment, and horizontal transfer has occurred, and where convergent or parallel evolution has played a role, the genetic relationships may not be strictly hierarchical (McCormack and Clewley, 2002). A network based on the split decomposition of the site-patterns doesn't force the data into the shape of the tree, and can, in fact, provide a good indication of how treelike the data are (Huson, 1998). We use a treelike topology to simulate our sequences, therefore, if there are conflicts present in the data, incorrect site-pattern recovery will lead to incorrect testing of the data.

In addition, a model that assumes rate homogeneity may result in inconsistency even if all other parameters of the model are correct (Gaut and Lewis, 1995). Gaut and Lewis simulated sequences under the K2P+Γ model of evolution using a Felsenstein tree with long branch lengths equal to 0.44 and short branch lengths equal to 0.04. The sequences were then analysed using an F85 model of evolution that does not take account of rate heterogeneity (i.e. assumes rate homogeneity). They estimated trees from sequences with lengths from 100 to 2500 nucleotides incremented in 100s. When the sequence lengths exceeded approximately 500 bp they observed that the maximum likelihood estimation was inconsistent leading to the reconstruction of the incorrect topology 100% of the time. Therefore, if rate heterogeneity is not accounted for in the model of estimation when it is present in the simulated model of evolution, our estimation is inconsistent. Similarly, if there is substantial rate variation in our biological sample, we will need to take account of this (Buckley and Cunningham, 2002; Sullivan and Swofford, 1997; Wakeley, 1994; Yang, 1996a). But, how often and how much rate heterogeneity is present in real biological data?

Rate heterogeneity is common in molecular sequences and the different regions along the DNA may be influenced by several factors. Protein coding DNA is the most extensively studied due to the functional importance of the proteins for which it codes. The different selective constraints at the different codon positions as well as in different regions of a gene lead to variable rates of nucleotide substitution. The same is true of ribosomal RNA genes as their gene products form secondary structures by pairing with themselves and have functional domains for binding ribosomal proteins. Therefore we expect that the complex nature of the constraints governing the nucleotide substitution process in these genes would lead to a high degree of among-site rate heterogeneity. This rate heterogeneity can be incorporated into the estimated

model of evolution by using a proportion of invariant sites, I (Hasegawa, Kishino and Yano, 1985), and/or a gamma parameter, Γ (Yang, 1994a), to model the distribution of rate heterogeneity. However, these are still not guaranteed to account for all observed rate heterogeneity and we may be required to partition the data and use more than one model of evolution on the data to accurately explain the observed heterogeneity (Bull et al., 1993).

4.1.3 Model Selection

The importance of the model in parametric simulation has been demonstrated in many places in the literature and is supported by work in this thesis. Due to the dependence of parametric simulation on the model, it is important when simulating the replicate data sets that we use the model that best fits the data. Objective model selection has been proposed to choose a model that avoids the dual problems of both underfitting and overfitting the model (Akaike, 1974; Burnham and Anderson, 1998; Goldman, 1993; Posada and Crandall, 1998). The relevance of model selection becomes apparent when the choice of different models of evolution changes the results of the analysis (Posada and Crandall, 2001c). Modeltest (v3.06 for PPC) is a program that chooses the model that best fits the data based on the likelihood scores generated from a PAUP* output (Posada and Crandall, 1998). The PAUP* procedure begins by constructing a neighbour-joining tree with a JC distance and then estimates the likelihood scores for 56 nested models of evolution. A neighbour-joining start tree is sufficient because it has been observed that the parameters of a model are robust with respect to the topology of the tree, provided that the tree is not chosen at random (Posada and Crandall, 2001a; Yang, 1994b; Yang, Goldman and Friday, 1994). The likelihood scores of nested models can be compared using Modeltest which performs

a series of hierarchical likelihood ratio tests (hLRTs). When models are nested (i.e. one is a special case of the other) the statistic is asymptotically distributed as χ^2 with q degrees of freedom, where q is the difference in the number of free parameters between the two models (Posada and Crandall, 2001a). Although the assumptions for the phylogeny problem may not strictly be met (Goldman, 1993), the hLRT seems to be robust against this (Yang, Goldman and Friday, 1995). The likelihoods for each model can also be compared using the Akaike Information Criterion, AIC (Akaike, 1974). The AIC is a useful measure that rewards models for good fit, but imposes a penalty for unnecessary parameters (Posada and Crandall, 1998). The key problem with increasing the number of parameters in the model is that the variance of the most crucial estimates becomes larger (Waddell, 1995). The effects of over-parameterising a model of evolution are not well understood and may under certain circumstances cause the model to develop unpredictable behaviour. In the case that a simpler model is determined statistically to have as good a fit to the data as a more complicated model, the simpler model is preferred (Burnham and Anderson, 1998). The parameters and the model chosen by Modeltest can then be used to build a more accurate tree, or in simulation studies such as this, simulate the replicate data sets.

4.2 Data

The data used in this chapter are empirical pre-aligned data sets for which there are conflicting evolutionary hypotheses from different data sources or analyses. In the case of the HIV-1, Primate, Rodent and Bird data sets there is a 'well-supported' topology for each based multiple other data sources including the fossil record, morphological analysis, Bayesian analysis, immunological data, various genetic data, and phylogenetic analyses of morphological characters. Although each of these has

been challenged at some stage, researchers generally agree that they are correct (Buckley, 2002; Buckley and Cunningham, 2002). Throughout this chapter any reference to a 'true' or 'correct' topology is a reference to the well-supported topology.

HIV-1 Data Set

Six homologous *gag* and *pol* sequences from four subtypes A (two sequences), B, D and E (two sequences) of 1996 bp in length are analysed for site-pattern frequencies. These sequences have previously been analysed Goldman et al. (2000) and Buckley (2002). However, four out of the original 2000 sites along the nucleotide alignment contained gaps and these have been stripped from the site-pattern analysis. The well-supported phylogeny has the two A sub-types together. However, when the *gag* and *pol* genes are concatenated, as in these data, it is known that there is sufficient conflict in the data to make A1 and A2 paraphyletic even under the most complex ML models.

Primate Data Set

This is the alignment of 888 bp of mitochondrial DNA (*mt*DNA) genes coding for transfer RNAs specific for histidine, serine and leucine from nine primate species. The sequences were published by Hayasaka et al. (1988) and are distributed with the PAML package. These data estimate the well-supported topology that is backed up by the fossil record, morphological analysis and Bayesian analysis.

Rodent Data Set

This data set contains *mt*12S rRNA and cytochrome B sequences from eight species of the genera *Peromyscus* (deermice) and *Onychomys* (grasshopper mice) sequenced

by Sullivan et al. (1995). The original 1078bp sequences had all positions with gaps removed to leave an alignment of 1066 nucleotides. When cytochrome B is analysed on its own, it will construct the well-supported topology. However, when the 12S rRNA is analysed on its own, it is known to construct a topology different from the well-supported topology even under maximum likelihood (Buckley, 2002; Sullivan, Holsinger and Simon, 1995). This concatenated data set constructs the well-supported topology.

Bee Data Set

This is a 677bp alignment of the mitochondrial cytochrome oxidase II region from six species in the genus *Apis* (honeybee). It is distributed as part of the SplitsTree package (Huson, 1998). No well-supported phylogeny has been determined for this data set, but SplitsTree analysis of these data show non-treelike conflicts in the data which make it interesting to site-pattern analysis.

Bird Data Set

This is an appropriately edited alignment of whole *mt*DNA genomes (14043bp) containing 36 of the 37 genes known to be constant across the vertebrates with alterations only in the gene order (Buckley, 2002; Buckley and Cunningham, 2002; Mindell et al., 1999). 24 of the genes code for a mature RNA product (22 tRNAs, the 23S rRNA and 16S rRNA). The remaining 12 genes code mRNAs are translated into proteins that are involved in oxidative phosphorylation while the 13th protein coding gene (ND6) is not included (Table 4.1). There are sequences from four bird species that represent certain key lineages and an alligator sequence (outgroup). Transition bias is particularly pronounced in animal mtDNA shown to be the case in birds by Edwards and Wilson (1990) All phylogenetic methods that have been used to analyse

this data set have led to the selection of an incorrect topology (Buckley and Cunningham, 2002; Mindell et al., 1999).

Full Name	Abbr.
NADH-ubiquinone oxidoreductase chain 6	ND6
NADH-ubiquinone oxidoreductase chain 5	ND5
NADH-ubiquinone oxidoreductase chain 4L	ND4L
NADH-ubiquinone oxidoreductase chain 4	ND4
NADH-ubiquinone oxidoreductase chain 3	ND3
NADH-ubiquinone oxidoreductase chain 2	ND2
NADH-ubiquinone oxidoreductase chain 1	ND1
Cytochrome c oxidase polypeptide III	COIII
Cytochrome c oxidase polypeptide II	COII
Cytochrome c oxidase polypeptide I	COI
Cytochrome b	CYT B
ATP synthase protein 8 (ATPase subunit 8)	ATP8
ATP synthase A chain (Protein 6)	ATP6

Table 4.1 The 13 protein products encoded by the mitochondrial genome.
ND6 is the only protein coding sequence that is not present in the bird whole genome *mt*DNA analysis The first 11 proteins are involved in the electron transport chain while the last two are involved in proton transport across the mitochondrial membrane.

4.3 Methods

4.3.1 Estimating the Model of Evolution

I used Modeltest for the selection of the model of evolution that best fits the data. Firstly, I constructed a NJ-tree for each of the data sets in PAUP*. Secondly, I estimated the likelihoods for 56 nested models of evolution on the NJ-tree. An example of these commands is supplied with Modeltest and the file to which the likelihood scores are saved is the input for Modeltest. Modeltest selects the model of evolution that best fits the data using hLRTs and the AIC. These criteria do not always agree on the model that best fits the data. Under these circumstances, I employed two different strategies. For the HIV-1 data set where two different models were chosen, I selected a composite model. To accommodate all parameters, the composite model was generalized to have both suggested models nested within it. For the Bee data set, two different models were also selected by the two criteria. In this

case, I estimated the parameters for both models and continued using both models for all subsequent analyses. An example of the PAUP command block for estimating the NJ-tree and parameters of the Bird data set under each of the 56 models of evolution is available in the appendix.

4.3.2 Simulating the Replicate Sequences

I simulated parametric replicates using Seq-Gen under the model of evolution selected by Modeltest and with the parameters and branch lengths for that model estimated by PAUP*. The command line entries for Seq-Gen are shown in Table 4.2. Three out of the six models suggested by Modeltest used the GTR model of nucleotide substitution. The remaining three selected models made restricted assumptions about the number of possible nucleotide substitution types in the rate matrix of the GTR model. In the matrix of the GTR model a-f are all able to vary, but if certain parameters are approximately equal, by making them equal to one another we can reduce the number of parameters that we are required to estimate, which results in increased speeds in estimation and a decrease in the variance. There are three rate matrix variants of the GTR model that resulted from using Modeltest on the five data sets. The TVM model (transversion model) uses the rate matrix (a b c d b e)[1] to have equal transitions, variable transversions and variable base frequencies. The TIM model (transition model) uses (a b c c e a)[1] to have equal transversions, variable transitions and variable base frequencies. Finally the K81uf model (Kimura, 1981; with unequal frequencies) uses (a b c c b a)[1].

---

[1] These letters refer to the restrictions on the possible substitutions in the r-matrix. See Section 1.6 Models of Evolution Figure 2.1 for the full schematic.

Data Set	Model	Seq-Gen Command Line
HIV-1	GTR+I+Γ	-n 1000 -l 1996 -m REV -f 0.39370812 0.16888945 0.21554975 0.22185267 -r 4.76515 19.89182 1.63861 2.71192 27.65441 1 -i
Primate	TVM+Γ	-n 1000 -l 888 -m REV -f 0.348557 0.314190 0.088005 0.249247 -r 11.36460 82.73557 7.76171 7.36028 82.73557 1.00000 -a 0.386827
Rodent	GTR+Γ	-n 1000 -l 1066 -m REV -f 0.36243390 0.22452362 0.15625359 0.25678888 -r 6.24819 13.12052 6.44792 4.08547 46.21096 1.00000 -a
Bees 1	K81uf+Γ	-n 1000 -l 677 -m REV-f 0.38449801 0.10926838 0.07618369 0.43004993 -r 1 16.56893 6.35928 6.35928 16.56893 1 -a 0.151750
Bees 2	TIM+I	-n 1000 -l 677 -m REV -f 0.38787003 0.09631150 0.08304275 0.43277572 -r 1 11.85762 6.93003 6.93003 30.80761 1 -i 0.70294333
Bird	GTR+I+Γ	-n 1000 -l 14043 -m REV -f 0.29839478 0.31091487 0.15111791 0.23957244 -r 4.29729 6.83106 1.43175 0.41624 7.68547 1 -a 0.600262 -i

Table 4.2 The Seq-Gen command lines for simulating the parametric replicates of each empirical data set.
The model selected by Modeltest and the parameters estimated using PAUP* were assembled as a Seq-Gen command line and used with the estimated tree to simulate replicate data sets that, in turn, were used to generate the null distribution of site-patterns.

4.3.3 Defining and Counting Site-Patterns

In any evolutionary tree, a branch partitions the tree into two connected components and consequently it partitions the set of taxa into two non-empty subsets, thus forming a split (Dress, Huson and Moulton, 1996). Site-patterns represent the splits at each site or column in the data on which the maximum likelihood tree is built (Page and Holmes, 1998). This analysis expands on the idea of a binary split to define the site-patterns with an intuitive process based on how the site-pattern data divides the taxa on a tree. Consider the fictional molecular sequence data shown in Table 4.3. The data can be summarised as observed site-patterns or data splits. We define the site-patterns using the following intuitive process. Firstly, we order the taxa in the data set, (in any order) and the taxa must be maintained in this order in the replicate data sets. Secondly, we assign the value '0' at every site for the first taxon in the data set. Any character state at a given site that has the same character state as taxon 1 is then also designated to have a '0' at that site. Thirdly, we assign '1' to the first site that has a character state at the given site that differs from Taxon 1. Every other taxon that has the same character state as this taxon is then also assigned '1'. Fourthly, we

assign '2' to the first site whose character state is different from both of these. Finally, we assign '3' if a taxon has a character state at the site that does not match any of the previous three. Since DNA consists of only 4 possible characters, 3 is the highest assignment that will ever be made to a character state at a given site. This is a similar principle to identifying parsimony informative sites, however, the limitations of "at least two groups with at least two taxa" for a parsimony informative site, does not apply. It is clear that in the case of the second site-pattern (Table 4.3), the grouping ((T1,T2,T3,T4)(T5)) is informative. It is also clear that the 10th site-pattern favours a grouping of T2 and T3. The only assumptions required for site-patterns of this description to be informative are that the sites are independent and identically distributed (Yang, 1996b). This site-pattern transformation is a form of data reduction that treats all nucleotides in the same manner.

Taxon 1	ACGT ACGT ACGT	0000 0000 0000
Taxon 2	ACCG ACGT AAGT	0011 0000 0100
Taxon 3	ACGT ACGG AAGG	0000 0001 0101
Taxon 4	ACTT CGGT AGGT	0020 1100 0200
Taxon 5	AGGC ACGG ATAT	0102 0001 0310

Table 4.3 The transformation process from DNA sequence to site-patterns.
12 sites from 5 fictional taxa are transformed from DNA sequence into site-patterns to illustrate the transformation process. The order of the data is not important when evaluating site-patterns, but the order that is chosen, must be consistent across the replicate data sets if comparisons are to be made.

To count these site-patterns I wrote a Perl script that simply followed the rules as laid out above (FinalHashDATASET[1]). It performs the site-pattern counting in an iterative fashion and can be displayed as a tab delimited data matrix in a spreadsheet. The first column is the observed site-pattern, the second column is the frequency of

[1] DATASET is one of Bird, Bees, HIV-1, Primate or Rodent. The full source code for FinalHashBird is available on the accompanying CD.

the site-patterns in the observed data and the following 1000 columns are the site-pattern frequencies for the 1000 parametric replicates. However, it is not often possible to display as a spreadsheet as there are far more rows and columns in the resulting data matrix than most software packages are able to handle. It is also not visually informative and so has not been included here.

4.3.4 Analysing Site-Pattern Data

Due to the limitations of the software on displaying and analysing data of this size, I wrote another Perl script to perform a meaningful analysis of the results (Gendist2.pl[1]). The simple stepwise process has three major parts.

(1) Find the average number of times that a given site-pattern was observed across all the replicate data sets.

(2) Create the S distribution for the expected variation across site-patterns from the individual values of S_j for each data set using the equation:

$$S_j = \sum_{i=1}^{n} \frac{(x_{ij} - \bar{x}_i)^2}{\bar{x}_i} \qquad \text{Eq. 4.1}$$

where x_{ij} is the frequency of site-pattern i for replicate data set j and $\bar{x}_i$ is average frequency across replicates for that site-pattern.

[1] Gendist2.pl is available on the accompanying CD.

(3) We then generate our test statistic, S^*, using the same equation as was used to calculate the site-pattern frequencies from the empirical data and test it against the parametric distribution of null values.

Using Eq. 4.1 we calculate the null values, S_j, for each of the replicates and the S^* test statistic for the observed data. These values give us a measure of the variation across the site-patterns. When one compares S^* to the null distribution for S, one can determine whether the site-pattern variation is typical or atypical of the distribution and discuss reasons for the good/poor performance of the parametric simulation method at recovering site-patterns in the data.

4.3.5 Investigating Non-treelike Evolution Using SplitsTree

Real data often contain conflicting phylogenetic signals and thus do not always clearly favour a unique tree (Huson, 1998). When data are ideal[1] SplitsTree gives rise to a tree. However, conflicting signals in the data are accounted for by the split decomposition method (Bandelt and Dress, 1992) implemented in SplitsTree. The taxa are reconstructed as a network rather than a strictly bifurcating tree. To investigate non-treelike influences, the Bee data set was used because there are conflicts in the data that are known to favour a network. The Bee data set was analysed by standard phylogenetic methods that assume a bifurcating tree in PAUP* and also using SplitsTree v2.4 for Macintosh (Huson, 1998) using the Jukes-Cantor model of evolution (Jukes and Cantor, 1969). A limitation of SplitsTree is that the parameters of the model cannot be determined from the program, so if we want to use

[1] ideal implies no more than to say there are no conflicts in the data.

those parameters to simulate sequences under the same model as was used to build the network, we must choose JC because then all the parameters of the rate matrix are set to one and the base frequencies are equal.

Parametric replicates were simulated under JC and the site-pattern recovery was analysed as described in Section 4.3.4. A comparison of the network produced by SplitsTree and the evidence for treelike evolution obtained from site-pattern analyses on the Bee data set allow us to answer questions like "Do we have treelike evolution? Is constraining our estimation procedure to be treelike informative or misleading? Are there perhaps some conflicts in our data that can arise through strictly bifurcating evolution that appear to support a non-treelike evolutionary hypothesis for our data?". If there are non-treelike influences in our data, simulating under a tree should violate that assumption and the replicate data sets will not be representative of the observed data.

4.4 Results

4.4.1 Selecting the Models and Parameters for the Data Sets

Models were selected for each of the five different data sets using Modeltest, which employs two selection criteria, a hLRTs and the AIC. For some of these data sets, the two different criteria select the same model, but in other data sets the suggested models are different. The HIV-1 data set was analysed and the models selected by Modeltest were not the same for the hLRTs and the AIC. TrN+Γ was selected as the best model by hLRTs and the GTR+I model was selected by the AIC. The TrN model (Tamura and Nei, 1993) makes some restrictive assumptions about the number of parameters in the GTR rate matrix. In the matrix of the GTR model a-f are able to

vary, but the varying parameters are limited to three in the matrix of the TrN (a b a a e a) where 'a' takes on a single value. More specifically, this allows the two transition types to have different rates, but all transversions have only a single rate. To accommodate both models, I generalised the model to a GTR+I+Γ model to include all the suggested substitution patterns of the matrix and both of the rate heterogeneity parameters I and Γ. The parameter values for the HIV-1 data set under the hLRTs suggested model (TrN+Γ) and the AIC suggested model (GTR+I) are shown in Table 4.4, as well as the parameters estimated by PAUP* for the GTR+I+Γ model used in the final estimation.

Parameter	hLRT suggested TrN+Γ	AIC suggested GTR+I	PAUP* Estimate
A	0.3960	0.3937	0.39370812
C	0.1747	0.1689	0.16888945
G	0.2133	0.2155	0.21554975
T	0.216	0.2219	0.22185267
a	1	4.7653	4.76515
b	7.864	19.8925	19.89182
c	1	1.6387	1.63861
d	1	2.7120	2.71192
e	10.9168	27.6553	27.65441
Γ	0.1932	None	Infinity
I	0	0.6719	0.67192

Table 4.4 The parameters estimated by PAUP* on the NJ-tree for the HIV-1 data set under the different models of evolution suggested by Modeltest.
We observe that the GTR+I estimates of the parameters on the neighbour-joining tree differ only after the third decimal place from the parameters of the tree constructed under maximum likelihood in PAUP* under GTR+I+Γ. This illustrates the robustness of estimating the parameters for models under a reasonable tree.

Another data set that presented a conflicting result after analysis with Modeltest was the Bee data set. Two different models of evolution were selected by the hLRTs and the AIC. The hLRTs selected the K81uf+Γ model of evolution and the AIC selected TIM+I. The parameters estimated by PAUP* on the NJ-tree are shown in Table 4.5. To see if this would make any difference to the recovery of the site-patterns, I estimated the parameters under both suggested models. Both had their parameters estimated in PAUP* using ML and both were used for the further site-pattern

analyses. The models of evolution that were selected by Modeltest for each data set that had their parameters estimated in PAUP* are shown in Table 4.6.

Parameter	hLRT suggested K81uf+G	AIC suggested TIM+I
A	0.3839	0.3871
C	0.1102	0.0971
G	0.0770	0.0840
T	0.4289	0.4318
a	1	1
b	15.3843	10.9962
c	6.5177	7.2168
d	6.5177	7.2168
e	15.3843	29.2615
Γ	0.1617	None
I	None	0.7007

Table 4.5 The parameters estimated by PAUP* on the NJ-tree for the Bee data set under the models of evolution suggested by Modeltest.
These parameters were compared in Modeltest by hLRTs and the AIC. The hLRTs suggested K81uf+G and the AIC suggested TIM+I. The parameters and model types seem sufficiently different to warrant separate estimation under each model.

Data Set	Selected Model
HIV-1	GTR+I+Γ
Primate	TVM+Γ
Rodent	GTR+Γ
Bees 1 (hLRT)	K81uf+Γ
Bees 2 (AIC)	TIM+I
Bird	GTR+I+Γ

Table 4.6 The models of evolution selected by Modeltest.
These models were specified for estimation of the tree and parameters for the data in PAUP*.

4.4.2 The Trees Reconstructed Under the Selected Model

The data sets chosen for analysis in this chapter do not all reconstruct the well-supported topology even when estimating with the most complex models of evolution (Table 4.7). The correct and incorrect topologies are displayed in Figure 4.1, with the correct tree placed above the incorrect tree.

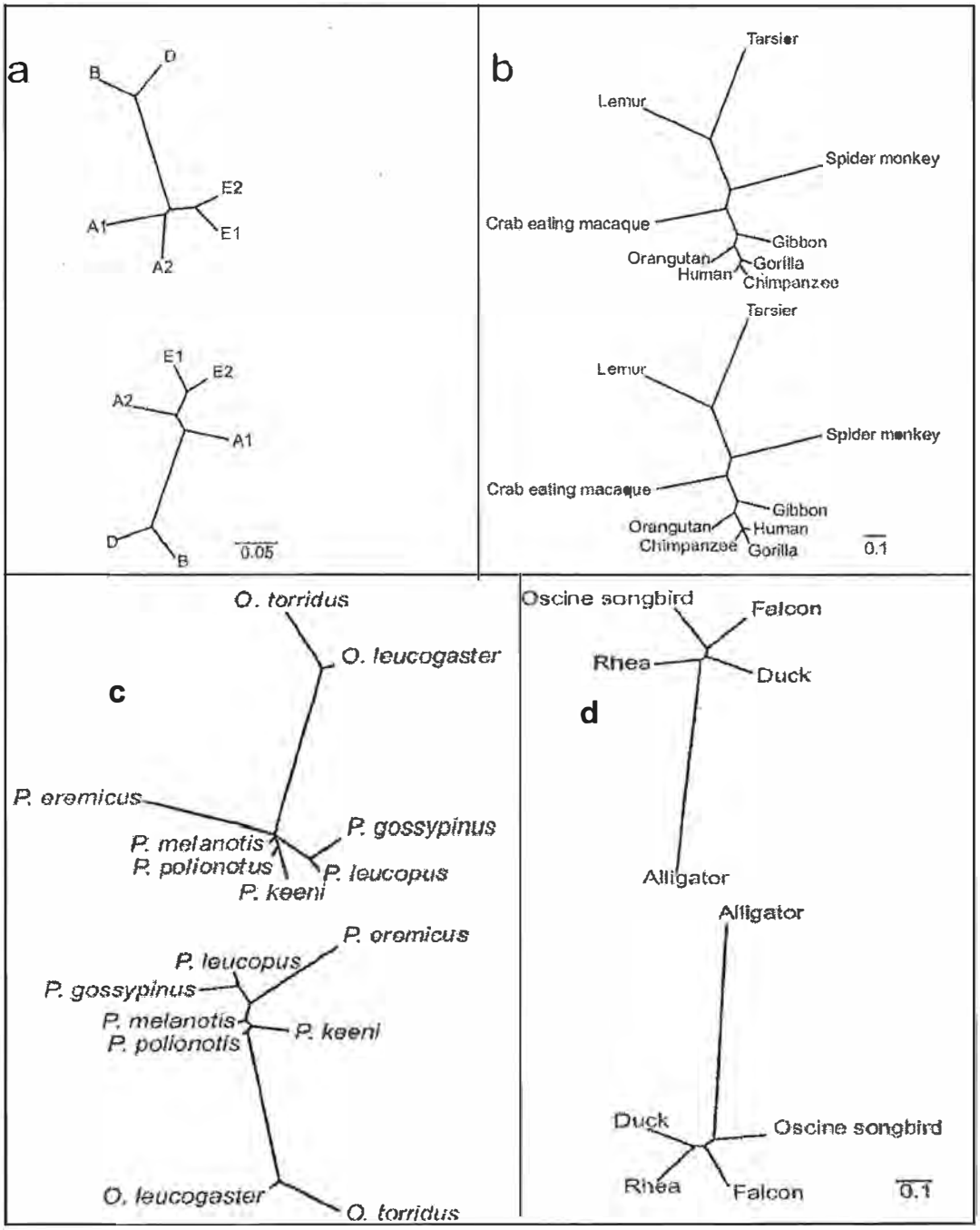

Figure 4.1 The correct[1] and incorrect topologies for each of the five data sets.
(a) The HIV-1 data set; (b) the Primate data set; (c) the Rodent data set; (d) the bird *mt*DNA for this data set. All images obtained from Buckley (2002).

[1] Incorrect or correct topologies are not guaranteed from empirical data. However, multiple independent data sources where the signal is unambiguous have given strong us strong evidence for certain topologies (Buckley and Cunningham, 2002; Buckley, 2002).

The Bird and HIV-1 data sets reconstruct the incorrect tree (Figure 4.1d,a). For the tree estimated from the Bird data set all the internal branches are relatively short while all the terminal branches are significantly longer. The support for the well-supported topology comes from multiple other sources including immunological data, various genetic data and phylogenetic analysis of morphological characteristics. In this data set the alligator sequence is the outgroup and its position elucidates the order of the speciation events. The well-supported hypothesis asserts that the ratites (represented here by the Rhea) were the earliest of the extant bird species to diverge. However, the tree estimated from whole *mt*DNA sequences indicates that the songbirds are the most basal. These contrasting hypotheses are indicative of a number of issues with using whole *mt*DNA as data for phylogenetic reconstruction. Amongst other things (to be mentioned in the discussion) it is typical of mitochondrial genomes to have very high levels of synonymous substitution, such that phylogenetic signal at the third codon position is saturated and therefore uninformative at high levels of divergence.

Data Set	Sequence Length	Number of genes[1]	Correct Topology
HIV-1	1996	2	NO
Primate	888	3	YES
Rodent	1066	2	YES
Bee	677	1	N/A
Bird	14043	36	NO

Table 4.7 Summary of data set information.
Not all the data sets construct the well-supported topologies. The sequences cover a range of sequence lengths and a varying number of genes.

The well-supported topology for the HIV-1 data set has A1 and A2 monophyletic. In Figure 4.1b we see that they are paraphyletic on the estimated tree. This is not as

[1] A gene is defined *sensu* Li (1997) as: A sequence of genomic DNA or RNA that performs a specific function. Performing the function may not require the gene to be translated or even transcribed.

significant a topological change as that observed in the Bird data set. However, it is still important because the process of determining subtype is a phylogenetic procedure. Both of these sequences had previously been determined through other phylogenetic analyses to be subtype A, but when *gag* and *pol* are concatenated they do not form a monophyletic subtype (i.e. the well-supported topology is not reconstructed).

The HIV-1 and Bird data sets have been chosen deliberately for this section because it is known that even under the models of best fit, the data are unable to reconstruct the 'correct' tree. Although they consist of a number of conflicting signals, the trees reconstructed from the other data sets do estimate the correct topology under models of evolution sufficiently complex to encompass the underlying nucleotide substitution process. Of further interest is the ability of the model that was used to estimate the trees, to recover the variation in the site-patterns of the original data.

4.4.3 Visual Analysis of the Raw Values

Visual analyses could not be applied to all data sets as the number of site-patterns and the number of replicates generated very large data matrices. However, we are able to analyse small portions of the data visually to give us an idea of what to expect from the statistical analyses. Table 4.8 shows the raw values of our analysis of the observed data and four replicate data sets from the Bird *mt*DNA site-pattern analysis. It is very unlikely that over the course of simulating 14043 nucleotide residues according to the best fit model of evolution we would recover the exact frequencies of the site-patterns in the observed sequence data. However, if the simulations are recovering the site-patterns correctly we would expect on average that the site-pattern frequencies would tend towards the true frequencies. Furthermore, if the difference between the

observed data and the mean is atypical of the difference between any of the replicate data sets and the mean, then we observe that the recovery of the site-patterns in the parametric procedure is poor. This may be due to multiple effects which will be discussed later.

Site-Pattern	Bird Data	Replicate 1	Replicate 2	Replicate 3	Replicate 4
'01023'	14	22	27	22	22
'00102'	147	196	173	198	188
'01121'	51	44	39	49	59
'01221'	48	37	44	38	40
'00012'	181	189	214	227	194
'00000'	7703	7866	7841	7898	7875
'01231'	13	8	9	12	7
'01001'	166	153	141	155	146
'01002'	139	196	193	186	193
'01100'	127	73	67	76	61

Table 4.8 A sample of the data matrix generated from site-pattern counting in the Bird data set.
This selection of 10 site-patterns from the observed bird *mt*DNA sequence data and four of the parametric replicate data sets shows some patterns with bigger relative differences than others. The numbers are the observed frequencies for the given site-pattern in each of the data sets.

The site-pattern '001100' is the 10^{th} site-pattern presented in Table 4.8. The true value from the data is 127 while all other values are substantially lower (73, 67, 76, and 61). This initial visual inspection illustrates potentially significant differences between the true values and the replicates. If this is typical of the site-pattern recovery for the data, then our test statistic will show that our result is atypical of the null distribution of site-patterns.

4.4.4 Null Distributions of Site-Pattern Variation

For each of the observed data sets in this chapter, the test statistic, S^*, was tested against its null distribution of 1000 replicates. In the HIV-1, Primate, Rodent and Bee data sets, S^* was non-significant (typical) when compared to the null distribution. However, for the Bird data set the test statistic S^* was very far removed from the null

distribution and therefore highly significant (atypical). Histograms for the S distribution including the placement of the S^* test statistic and the $\alpha = 0.05$ significance value are shown in Figure 4.2 for all of these data sets.

The results of the comparison of each test statistic to its null distribution are summarised in Table 4.9. The Bird data set is almost the entire mitochondrial genome consisting of 36 out of 37 genes. These same 37 genes are present in every vertebrate animals studied to date, and vary only in their relative position (Macey et al., 1997). The genes code for 22 tRNAs 2 rRNAs and 13 proteins. Each of these has different substitution rates which allows for suitable regions to be chosen for the question under study. However, by concatenating a number of genes as a single molecular sequence, each of which undoubtedly has a different underlying model of evolution, the model is again misspecifiied.

Dataset	Well-Supported Topology	Shape of Distribution	$\alpha = 0.05$	S^*	Significant
Bee 1	Yes	Right Skew	481	86.4	No
Bee 2	Yes	Right Skew	581	54	No
Bird	No	Symmetric	68.7	297	Yes
HIV	No	Right Skew	317.7	240	No
Primate	Yes	Symmetric	11629	5136	No
Rodent	Yes	Right Skew	2274	177	No

Table 4.9 The characteristics of each null distribution.
A significant result is one where the site-pattern variation in the observed data is atypical of the null distribution, because S^* is greater than the 95% significance level of the distribution.

Although the HIV-1 data set is the concatenation of two different genes from the viral genome we observe that even though the incorrect tree is estimated from the data and used to simulate the replicate data sets, the site-pattern analysis does not reject the null hypothesis. The *gag* gene codes for internal virion proteins and *pol* gene codes for the enzymes reverse-transcriptase, protease and integrase. Since their gene products perform different functions they are most likely under different evolutionary constraints. In fact it is known that *gag* has a markedly higher mutation rate than *pol*

(McCormack and Clewley, 2002) and both are less variable than *env*, the most commonly studied HIV-1 gene. Therefore, the underlying model for each of *gag* and *pol* within the combined sequence will be different. By concatenating the two genes and attempting to estimate a single model, we induce an error in the model that manifests in the incorrect reconstruction of the phylogeny. However, even when the model was deliberately misspecified and T_{ML} was used instead of T_0, it was not enough to significantly affect the site-pattern recovery of the HIV-1 data.

In the other data sets, where the model is well specified for the data, the site-pattern recovery is very accurate. In the Rodent data set, where similarity between sequences is high and the correct topology is obtained during estimation, only 13 replicate data sets showed less variation from the mean site-pattern frequencies than the observed data. This is a result mimicked in the Primate data set where only 10 replicate data sets showed less variation in the site-patterns analyses than the observed data. In these situations we would have a strong degree of confidence in the accuracy of the model for estimating the subsequent phylogenetic tree and inferring the evolutionary relationship among the taxa. Furthermore, for the Bee data set where two different models were specified, the recovery of the site-pattern data remains typical of the null distribution regardless of the model that is used.

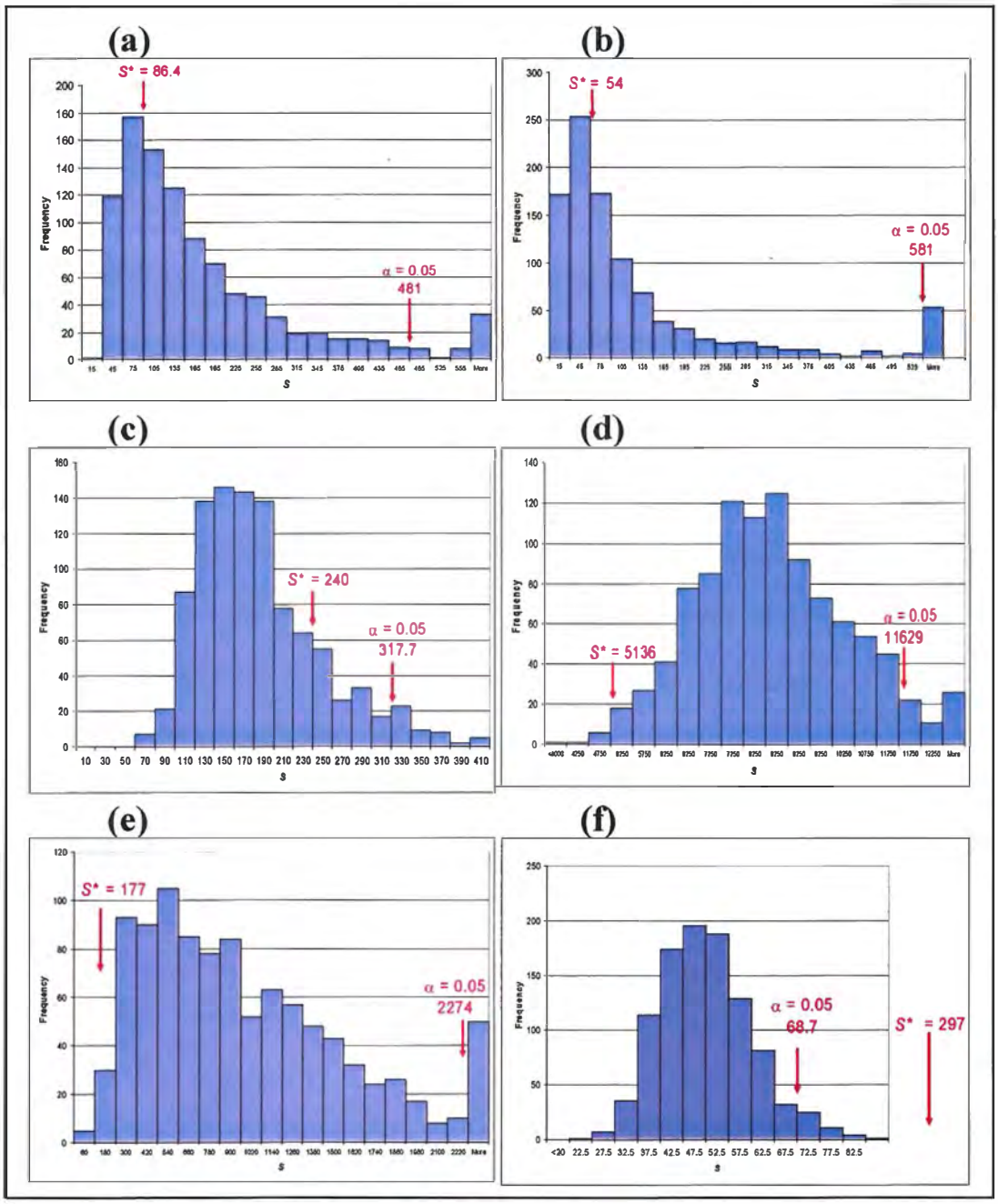

Figure 4.2 The Null distribution for the site-pattern variation in all data sets.
These are the null distributions of (a) the Bee data set simulated under the K81uf+G model of evolution; (b) the Bee data set simulated under the TIM+I model of evolution; (c) the HIV-1 data set simulated under the GTR+I+Γ model of evolution; (d) the Primate data set simulated under the TVM+Γ model of evolution; (e) the Rodent data set simulated under the GTR+Γ model of evolution; (f) the Bird data set simulated under the GTR+I+Γ model of evolution.

4.4.5 The Bee Data Set: Tree or Network?

The Bee data set was analysed using SplitsTree with the JC model of evolution to produce Figure 4.4. If the data are non-treelike we see a network of the relationship between the taxa on the tree. However, if the data are strictly treelike we see that the data conform to a tree. Analysing the data in this way will allow us to gain a measure of how treelike or network-like the data are. For the Bee data set SplitsTree generates a network Figure 4.4, which is evidence against the hypothesis of diverging or treelike evolution.

The correct interpretation of this network is to say that there are conflicting signals in the data that induce the network pattern observed. It is most prominent between *Apis mellifer, A. dorsata* and *A. cerana.* From this we are unable to determine to which taxon *A. mellifer* is most closely related. When the Bee data set was estimated using PAUP* on the model of best fit selected by Modeltest, *A. mellifer* is grouped most closely with *A. cerana* (Figure 4.3). Which of these two interpretations of the same data set is more accurate?

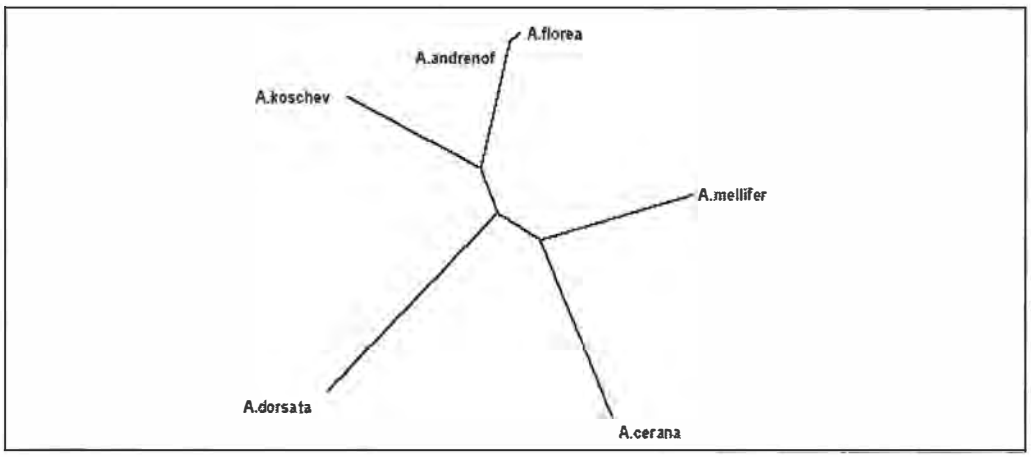

Figure 4.3 The ML tree for the Bee data set under the JC model of evolution.

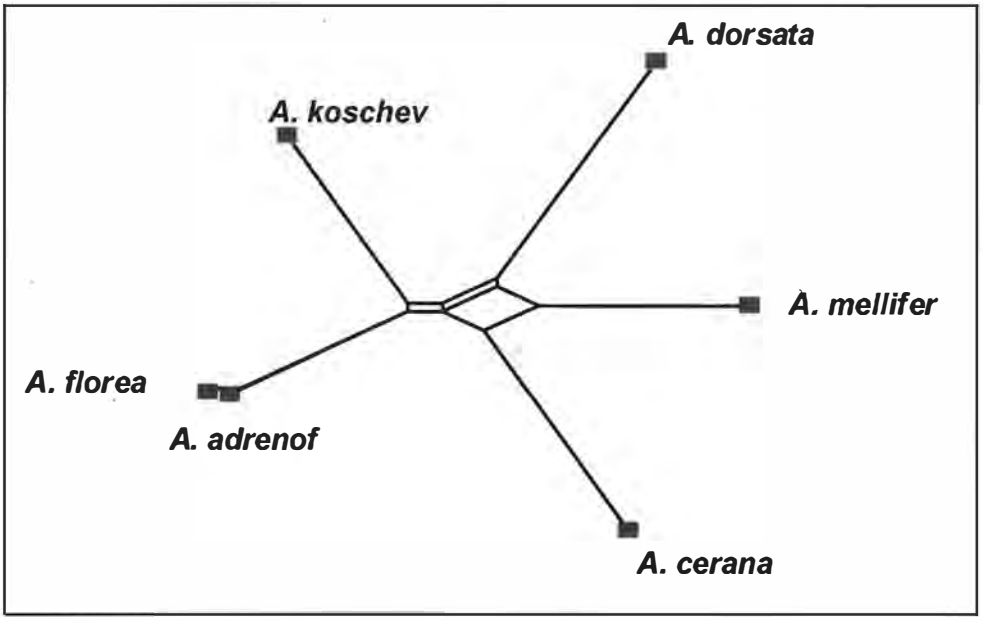

Figure 4.4 The network reconstruction of the Bee data set under the JC model in SplitsTree.
The position of *A. **mellifera*** shows the most conflict and using this method its closest related taxon cannot be determined.

For example, we can suggest as the null hypothesis for the Bee data set, that the data are treelike (i.e. our sampled taxa did not undergo recombination, reassortment or horizontal gene transfer). I simulated the sequences under the same model of evolution as was used to generate the network in SplitsTree (i.e. Jukes Cantor) using a tree to create the replicate data sets. If simulating under a tree can accurately recover the site-patterns under the same model of evolution as SplitsTree uses to suggest that a network is a better explanation of the data, then we cannot reject the null hypothesis that the data are treelike. The results of the site-pattern analysis assuming a JC model of evolution gives us the S distribution and S^* statistic as shown in Figure 4.5. The difference between the results that are obtained through analysis with SplitsTree and the results obtained through parametric simulation of site-patterns are contradictory. These data suggest that the conflicting splits which SplitsTree interprets as non-treelike evolution can simply arise through treelike evolution.

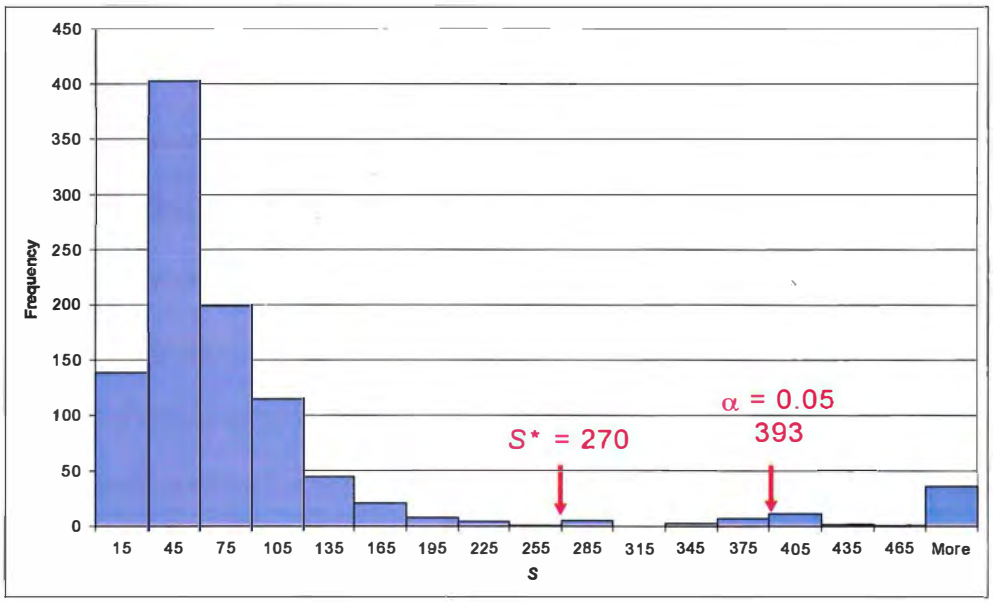

Figure 4.5 Null distribution for the site-pattern variation in the Bee data set under JC. These data show that we are unable to reject the JC model of evolution and that the evolution of the original sequences is treelike because the S^* statistic is typical of the null distribution.

4.5 Discussion

4.5.1 Site-Pattern Recovery

The results presented here suggest that parametric bootstrapping accurately recovers site-patterns most of the time for these data sets and that the accuracy appears to be data set dependent. Site-pattern recovery is influenced by two main elements. The first is non-treelike artefacts in the data. These would invalidate the use of a start tree to simulate the replicate data sets. The second is the adequacy of the model of evolution estimated from the data. The importance that selecting the model has on the performance of simulating data sets under a Monte Carlo method is already known from work in this thesis and in many places in the literature (Buckley, 2002; Goldman, 1993; Goldman, Anderson and Rodrigo, 2000; Huelsenbeck, Hillis and Nielsen, 1996). In addition, a distinction needs to be made clear on the difference between model selection (or choice) and model adequacy (or assessment). Model

selection is a relative measure that implements a criterion to decide on which of the available models has performed the best at explaining the data (Burnham and Anderson, 1998). Model adequacy is an absolute measure of how well the model fits the data (Bollback, 2002). Essentially the selected model of best fit may still be inadequate at explaining the data. In fact, most models appear to be poor descriptions of sequence evolution (Goldman, 1993). While this may generally be true, our methods are often robust to violations in the assumptions that are required to describe sequence evolution (Sullivan and Swofford, 2001). Therefore, using the best available methods and making as few assumptions as possible while being as biologically realistic as possible is of quite some importance. Modeltest is, in fact, very efficient at selecting the model that best fits the data (Posada and Crandall, 2001a), but it does not give any information on how good the fit is. This can only be determined through other testing strategies. If a single model does not describe the data adequately Modeltest is not capable of giving us the insight we require. As a prudent investigator, one would usually have some idea of the characteristics of the DNA sequence we are investigating, but the extent to which these artefacts may affect an analysis needs to be determined directly from the data we have.

4.5.2 Testing Treelike Evolution

The null hypothesis that the evolution in the Bee data set is treelike was unable to be rejected by the site-pattern analysis even though SplitsTree presents the data as a network. In the absence of further evidence, using a tree to represent the data is therefore adequate for the Bee data set. In addition, it follows that the conflicts in the data that appear to be evidence for network evolution can arise naturally through stochastic treelike evolution.

It is more intriguing to note that the hypothesis of treelike evolution is still unable to be rejected even though the model of evolution used to simulate the sequences is misspecified (JC and not the model of best fit). This provides two possible explanations. One possibility is that SplitsTree is incorrectly interpreting conflicts in the data as non-treelike artefacts. Alternatively, our site-pattern test may be too conservative and therefore unable to reject the null hypothesis when it is false. Future work may involve integrating over the uncertainty in the branch lengths for the start tree and take n random draws from the marginal distributions to see if this makes any difference to the sensitivity of the site-pattern test. However, it may be that for site-pattern tests to have significant application in model selection or testing treelike evolution, we require more power in the test statistic. Numerous test statistics can be formulated, but to be useful they have to represent a relevant summary of the model parameters and the data (Bollback, 2002). Modifications to the test statistic that achieve the desired power to reject the null-hypothesis when it is false will require further investigation. Comparison with this and other methods will be required to fully justify its application, and is certainly an appealing avenue for further research. One way in which we can test the sensitivity of new test statistics is to simulate the replicate data sets using trees that have a topology different from the well-supported topology, from data sets that actually do estimate the well-supported topology. A similar approach could be taken in a simulation study in which the true tree is known to provide us with more certainty in our method. In these situations, we would expect could assess the type I and type II errors and compare them to provide a recommendation of the best available method.

4.5.3 Is a Single Model Sufficient?

It is quite clear that selective pressures act to different degrees and in different ways on nucleotides at each of the three codon positions. Nucleotide substitutions that lead to synonymous (silent) substitutions are significantly more frequent that those that lead to non-synonymous (amino acid altering) substitutions (Li, 1997). Therefore it is clear that the 3rd codon position is able to vary much more freely than either the first or second, because changes here lead more frequently to synonymous substitutions than do changes at the other positions. Yang (1996c) analysed sequences from six hominoid species estimating the rates of change in the three codon positions separately and showed that the second position changed at a rate 3 times slower than that of the first position and that the third position changed at over 20 times faster than the second position. This conclusion along with evidence from other studies (Sullivan and Swofford, 1997; Wakeley, 1994; Yang, Goldman and Friday, 1994; Yang, Goldman and Friday, 1995) show that ignoring rate variation among sites leads to model violation and causes underestimation of branch lengths.

The topologies found in tree searching for the first four data sets are not all consistent with the current evolutionary hypotheses even though rate heterogeneity has been accounted for. Although those phylogenetic relationships that have been proposed as 'correct' have been challenged at some stage, their topologies are supported by multiple independent sources (Buckley, 2002; Buckley and Cunningham, 2002). Therefore, in the case of the HIV-1 and the Bird data set, why are these data not also supporting the assumed correct topologies? In each of these data sets it is well understood that there are different models of evolution underlying different regions of the DNA. Interestingly in the case of the HIV-1 data set this only causes an incorrect

phylogeny to be reconstructed while the site-pattern analysis does not lead to rejection of the model that was estimated from the data. The phylogenetic reconstruction uses a model that takes account of specific nucleotides, the frequency of occurrences in the data, and their propensity to change to other nucleotides. However, the site-pattern transformation is a data reduction that treats all nucleotides in the same manner and looks only at the pattern of partitions that are produced. This reduction in the data may have contributed to the observed reduction in sensitivity or explanatory power of the analysis leading to it being a conservative test.

The bird data set obviously has more serious model misspecification as shown by the estimation of the incorrect topology and a rejection of the model under the parametric site-pattern analysis. Buckley (pers. comm.) commenting on the Bird data set:

> *The Bird data set has all the tRNA, rRNA and protein coding mtDNA genes. It includes 3rd positions so is fairly well saturated. There are differences in base frequencies among lineages so all models should be* ***badly violated****. For the Bird data set the well-supported tree is never recovered from these data.*

It is not surprising that when the model is badly violated and we cannot accurately infer the correct tree that the site-pattern recovery is aberrant. Even the inclusion of rate heterogeneity parameters to allow for multiple rates is not always sufficient as shown by this thesis and in the analyses by Buckley (2002). Furthermore, it does not seem likely that there is a single best fit model of evolution appropriate for most data sets (Muse, 1999). In the case of the Bird data set we see that the inability to accurately estimate a single model from this data set is mirrored in the performance of the parametric bootstrap to recover the site-patterns that were present in the original data. Therefore, it appears that a single model cannot accurately recover the site-patterns observed in these sequences.

4.5.4 Combining and/or Partitioning Data

Partitioning and combining data provides a point of contention in phylogenetic analysis with three strategies having been proposed.

(1) Kluge (1989) proposed that phylogenetic analysis should always be performed using all the data (total evidence). This was based largely on philosophical arguments where the 'total evidence solution' is sought because it maximizes the 'informativeness' and 'explanatory power' of the data. In practice this is only true when the method used in consistent.

(2) Miyamoto and Fitch (1995) took an alternative approach that suggested that the data should never be combined. The idea is that each partition can separately estimate a phylogeny, and these independent estimates can then be used to corroborate certain groupings of taxa

(3) Conditional combination based on a test of homogeneity (Bull et al., 1993; De Quieroz, 1993; Rodrigo et al., 1993) has been suggested as a third option (Huelsenbeck, Bull and Cunningham, 1996). Heterogeneous partitions are those which give rise to significantly different trees and should not be combined, otherwise combination is advocated.

Combining genes for analysis can be expected to lead to more reliable estimation of the phylogenetic relationships and more accurate calculation of branching dates (Yang, 1996c). However, Yang also concedes that different genes perform different functions and may have followed different evolutionary processes leading to a need for a different model for each gene leading to methods becoming inconsistent. Bull et al. (1993) discuss the merits of partitioning and combining in some depth and

advocate the prudent application of a test before deciding. The most fundamental message is that the level of heterogeneity must be evaluated before proceeding with a partitioned or combined analysis. Effectively the problem reduces to a standard statistical problem of analysis of variance such that heterogeneity must first be rejected before data can be combined. If heterogeneity cannot be rejected (i.e. the data are demonstrably heterogeneous) they should not be combined unless the analysis can appropriately take account of this heterogeneity.

Partition homogeneity tests are not new [e.g. incongruence length difference (ILD) test, the bootstrap-support test proposed by De Queiro (1993), the distance between minimum length trees (Rodrigo et al., 1993) and the LRT of conflicting phylogenetic signal (Huelsenbeck and Bull, 1996)]. If the data are not incongruent (i.e. estimate the same topology), supporters of 'conditional combination' suggest that combining the data would improve the phylogenetic accuracy over individual analyses of the partitions. The test suggested by Huelsenbeck and Bull (1996) proposes the null hypothesis that the same topology underlies each gene or each gene region under study. The alternative hypothesis is that different trees and different evolutionary rates can underlie each gene or gene region. A standard likelihood ratio test can be applied in this situation because H_0 is nested within H_1. For example, if we analyse the HIV-1 data set which we know to be a concatenation of two different genes, it would certainly be beneficial to partition the data into the two regions and estimate the maximum likelihood for each region independently. A simple LRT would suffice to compare the likelihoods for the single model and for the 2 model system. If the test is significant, we determine that the single model is not sufficient to explain the concatenated data set (e.g. the HIV-1 data set that is a concatenation of *gag* and *pol*) and that a two model system should be used to explain the data and build the

phylogenetic tree. A similar approach could be used on the Bird data set, however, with the 36 genes in this data set we may have a large number of partitions.

However, a significant problem associated with partitioning remains. We may not always have the luxury of working with data where the boundaries of the partitions are well defined or already known. For example, if we have a gene that has a poor fit to a single model, we will need to partition the data ourselves. This action would be highly subjective as the number of partitions and relative sizes of the partitions would likely be variable between different investigators. Another approach that would minimize the subjectivity that could be used when a single model does not appear to fit the data adequately, would be to create a sliding window of specified length and slide it along the genome calculating the likelihoods and model parameters for each region. The size of the window and the width of the slide would still be a user-defined choice, but through trying a number of options a degree of thoroughness and robustness can be produced. We can then generate a LRT that compares two likelihoods for the region in the window. The likelihood of the data given the tree under the model and parameters estimated from the entire sequence can be compared against the likelihood of the best fit model for the sliding window. If there is a statistical difference between the two, then it is clear that the overall model, calculated from the entire genome, is not sufficient to model the region in the sliding window. The best way to continue is to determine biologically why this region (or multiple regions) does not fit the model and partition the data with biological relevance to model each partition separately. This approach would be quite sensitive to the changes in the data and should detect any significant changes in the model. Furthermore, we will also need to consider that the data may need to be partitioned based on codon position. Here a sliding window approach would not be appropriate,

but a method similar to Yang et al. (1996c) would serve a similar purpose in determining whether or not multiple models should be incorporated based on codon position.

5 Final Conclusion

Parts of this thesis followed reasonably well defined methods already implemented and working correctly. However, as is the norm with research, novel results lead to the need for new or modified procedures. Taking the old and forging the new in the context of this thesis has opened up a number of avenues that can, in time, be better understood, but not without more rigorous scientific work. The main results from this thesis are:

- Evidence that agrees with Shibata's principle that overfitting models is almost certainly better than underfitting models (Shibata, 1989).

- The results in Chapter 2 agree with the ideas of Goldman, Anderson and Rodrigo (2000) and Aris-Brosou (2003a; 2003b) that the different form of the simple null hypothesis in topology testing accounts for a significant portion in the difference in performance observed between different tests.

- The Results in Chapters 2 and 3 corroborate the ideas of Huelsenbeck, Hillis and Nielsen (1996) that insufficiently complex models of evolution lead to excessive type I error. Therefore, it seems reasonable to recommend that the most complex models available should always be used.

- The methods described in Chapter 4 provides a novel exploration of data and an alternative way of performing model selection and testing model adequacy.

It is highly evident that this is merely a launching point for further analyses and investigation into the testing of phylogenetic hypotheses. The use and accuracy of models of evolution to infer phylogenies from sequences is still hotly debated and with the seemingly limitless data sources available I envisage that time and effort integrating new techniques with new data will lead us rapidly towards a deeper understanding.

6 References

Akaike, H. 1974. A New Look at the Statistical Model Identification. *IEEE Trans. Autom. Contr.* **19**:716-723.

Antezana, M. 2003. When Being "Most Likely" is Not Enough: Examining the Performance of Three Uses of the Parametric Bootstrap in Phylogenetics. *J. Mol. Evol.* **56**:198-222.

Aris-Brosou, S. 2003. Least and Most Powerful Phylogenetic Tests to Elucidate the Origin of the Seed Plants in the Presence of Conflicting Signals under Misspecified Models. *Syst. Biol.* **52**:781-793.

Aris-Brosou, S. 2003a. How Bayes Tests of Molecular Phylogenies Compare with Frequentist Approaches. *Bioinformatics* **19**:618-624.

Aris-Brosou, S. 2003b. Least and Most Powerful Phylogenetic Tests to Elucidate the Origin of the Seed Plants in the Presence of Conflicting Signals under Misspecified Models. *Syst. Biol.* **52**:781-793.

Bandelt, H.-J., and A. Dress. 1992. Split Decomposition: A New and Useful Approach to Phylogenetic Analysis of Distance Data. *Mol. Phylogenet. Evol.* **1**:242-252.

Bollback, J. 2002. Bayesian Model Adequacy and Choice in Phylogenetics. *Mol. Biol. Evol.* **19**:1171-1180.

Box, G. 1976. Science and Statistics. *Journal of the American Statistical Association* **71**:791-799.

Buckley, T. 2002. Model Misspecification and Probabilistic Tests of Topology: Evidence from Empirical Data Sets. *Syst. Biol.* **51**:509-523.

Buckley, T., and C. Cunningham. 2002. The Effects of Nucleotide Substitution Model Assumptions on Estimates of Nonparametric Bootstrap Support. *Mol. Biol. Evol.* **19**:394-405.

Bull, J., J. Huelsenbeck, C. Cunningham, D. Swofford, and P. Waddell. 1993. Partitioning and Combining Data in Phylogenetic Analysis. *Syst. Biol.* **42**:384-397.

Burnham, K., and D. Anderson. 1998. Model Selection and Inference: A Practical Information-theoretic Approach. Springer Verlag, New York.

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

Cavender, J. 1978. Taxonomy with Confidence. *Math. Biosci.* **40**:270-280.

Chernoff, H., and L. Moses. 1959. Elementary Decision Making. John Wiley and Sons, New York.

Cox, D., and H. Miller. 1977. The Theory of Stochastic Processes. Chapman and Hall, London.

Cummings, M. 2003. Comparing Bootstrap and Posterior Probaility Values in the Four-Taxon Case. *Syst. Biol.* **52**:447-487.

De Quieroz, A. 1993. For consensus (sometimes). *Syst. Biol.* **26**:657-681.

De Quieroz, K., and S. Poe. 2003. Failed Refutations: Further Comments on Parsimony and Likelihood Methods and their Relationship to Popper's Degree of Corroboration. *Syst. Biol.* **52**:352-367.

Dress, A., D. Huson, and V. Moulton. 1996. Analyzing and Visualizing Sequence and Distance Data Using SplitsTree. *Discrete Applied Mathematics and Combinatorial Operations Research and Computer Science* **71**.

Eck, R., and M. Dayhoff. 1966. Atlas of Protein Structure 1966. National Biomedical Research Foundation, Silver Spring, Maryland.

Edwards, A., and L. Cavalli-Sforza. 1963. The Reconstruction of Evolution. *Heredity* **18**:553.

Edwards, S., and A. Wilson. 1990. Phylogenetically Informative Length Polymorphisms and Sequence Variability in Mitochondrial DNA of Australian Songbirds. *Genetics* **126**:695-711.

Efron, B. 1985. Bootstrap Confidence Intervals for a class of Parametric Problems. *Biometrika* **72**:45-58.

Efron, B., E. Halloran, and S. Holmes. 1996. Bootstrap Confidence Levels for Phylogenetic Trees. *Proc. Natl. Acad. Sci. USA* **93**:13429-13434.

Felsenstein, J. 1973. Maximum Likelihood Estimation of Evolutionary Trees from Continuous Characters. *American Journal of Genetics* **25**:471-492.

Felsenstein, J. 1978a. The Number of Evolutionary Trees. *Syst. Zool.* **27**:27-33.

Felsenstein, J. 1978b. Cases in which Parsimony or Compatibility Methods Will be Positively Misleading. *Syst. Zool.* **27**:401-410.

Felsenstein, J. 1981a. Evolutionary Trees from DNA Sequences: a Maximum Likelihood Approach. *J. Mol. Evol.* **17**:368-376.

Felsenstein, J. 1981b. Evolutionary Trees from Gene Frequencies and Quantitative Characters: Finding Maximum Likelihood Estimates. *Evolution* **35**:1229-1242.

Felsenstein, J. 1984. Distance Methods for Inferring Phylogenies: A Justification. *Evolution* **38**:16-24.

Felsenstein, J. 1985. Confidence Limits on Phylogenies: An Approach Using the Bootstrap. *Evolution* **39**:783-791.

Felsenstein, J. 1988. Phylogenies From Molecular Sequences: Inference and Reliability. *Annu. Rev. Genet.* **22**:521-565.

Gaut, B., and P. Lewis. 1995. Success of Maximum Likelihood Phylogeny Inference in the Four-Taxon case. *Mol. Biol. Evol.* **12**:152-162.

Goldman, N. 1993. Statistical Tests of Models of DNA substitution. *J. Mol. Evol.* **36**:182-198.

Goldman, N., J. Anderson, and A. Rodrigo. 2000. Likelihood-based Tests of Topologies in Phylogenetics. *Syst. Biol.* **49**:652-670.

Guindon, S., and O. Gascuel. 2003. A Simple, Fast and Accurate Algorithm to Estimate Large Phylogenies by Maximum Likelihood. *Syst. Biol.* **52**:696-704.

Hacking, I. 1965. Logic of Statistical Inference. Cambridge University Press.

Hall, P., and M. Martin. 1988. On Bootstrap Resampling and Iteration. *Biometrika* **75**:661-671.

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.

Hayasaka, K., T. Gojobori, and S. Horai. 1988. Molecular Phylogeny and Evolution of Primate Mitochondrial DNA. *Mol. Biol. Evol.* **5**:626-644.

Hendy, M., and D. Penny. 1982. Branch and Bound Algorithms to Determine Minimal Evolutionary Trees. *Math. Biosci.* **59**:277-290.

Hillis, D., B. Mable, and C. Moritz. 1996. Application of Molecular Systematics: State of the Field and a Look to the Future. Pages 515-543 *in* Molecular Systematics (D. Hillis, B. Mable, and C. Moritz, eds.). Sinauer, Sunderland, Massachusetts.

Huelsenbeck, J. 1995. The Robustness of Two Phylogenetic Methods: Four-Taxon Simulations Reveal a Slight Superiotiry of Maximum Likelihood over Neighbor-Joining. *Mol. Biol. Evol.* **12**.

Huelsenbeck, J., and J. Bull. 1996. A Likelihood Ratio Test to Detect Conflicting Phylogenetic Signal. *Syst. Biol.* **45**:92-98.

Huelsenbeck, J., J. Bull, and C. Cunningham. 1996. Combining Data in Phylogenetic Analysis. *TRENDS in Ecology and Evolution* **11**:152-158.

Huelsenbeck, J., and K. Crandall. 1997. Phylogeny Estimation and Hypothesis Testing Using Maximum Likelihood. *Annu. Rev. Ecol. Syst.* **28**:437-66.

Huelsenbeck, J., D. Hillis, and R. Nielsen. 1996. A Likelihood Ratio Test of Monophyly. *Syst. Biol.* **45**:546-558.

Huelsenbeck, J., and R. Nielsen. 1997. The Effect of non-Independent Substitution on Phylogenetic Accuracy. *Syst. Biol.* **48**:317-328.

Huson, D. 1998. SplitsTree: a Program for Analysing and Visualising Evolutionary Data. *Bioinformatics* **14**:68-73.

Jukes, T., and C. Cantor. 1969. Evolution of Protein Molecules. Pages 21-132 *in* Mammalian Protein Metabolism (H. Munro, ed.) Academic Press, New York.

Kimura, M. 1980. A Simple Method for Estimating Evolutionary Rate of Base Substitutions Through Comparative Studies of Nucleotide Sequences. *J. Mol. Evol.* **16**:111-120.

Kimura, M. 1981. Estimation of Evolutionary Distances Between Homologous Nucleotide Sequences. *Proc. Natl. Acad. Sci. USA* **78**:454-458.

Kishino, H., and M. Hasegawa. 1989. Evaluation of the Maximum Likelihood Estimate of the Evolutionary Tree Topologies from DNA Sequence Data, and the Branching Order in Hominoidea. *J. Mol. Evol.* **29**:170-179.

Kluge, A. 1989. A Concern for Evidence and a Phylogenetic Hypothesis of Relationships Among Epicrates. *Syst. Zool.* **38**:7-25.

Kuhner, M., and J. Felsenstein. 1994. A Simulation Comparison of Phylogeny Algorithms under Equal and Unequal Evolutionary Rates. *Mol. Biol. Evol.* **11**:459-468.

Lanave, C., G. Preparata, C. Saccone, and G. Serio. 1984. A New Method for Calculating Evolutionary Substitution Rates. *J. Mol. Evol.* **20**:86-93.

Li, W.-H. 1997. Molecular Evolution. Sinauer Associates, Sunderland, MA.

Macey, J., A. Larson, N. Ananjeva, Z. Fang, and T. Papenfuss. 1997. Two Novel Gene Orders and the Role of Light-Strand Replication in Rearrangement of the Vertebrate Mitochondrial Genome. *Mol. Biol. Evol.* **14**:91-104.

Maddison, D., and W. Maddison. 2000. MacClade 4.0, version 4.0. Sinauer Associates Inc.

McCormack, P., and J. Clewley. 2002. The Application of Molecular Phylogenetics to the Analysis of Viral Genome Diversity and Evolution. *Reviews in Medical Virology* **12**:221-238.

Mindell, D., M. Sorenson, D. Dimcheff, M. Hasegawa, J. Ast, and T. Yuri. 1999. Interordinal Relationship of Birds and Other Reptiles Based on Whole Mitochondrial Genomes. *Syst. Biol.* **48**:138-152.

Miyamoto, M., and W. Fitch. 1995. Testing Species Phylogenies and Phylogenetic Methods with Congruence. *Syst. Biol.* **44**:64-76.

Muse, S. 1999. Modeling the Molecular Evolution of HIV Sequences. John Hopkins University Press, Baltimore, Md.

Neyman, J. 1950. First Course in Probability and Statistics. Henry Holt, New York.

Page, R. 1996. TREEVIEW: An Application to Display Phylogenetic Trees on Personal Computers. *Computer Applications in the Biological Sciences* **12**:357-358.

Page, R., and E. Holmes. 1998. Molecular Evolution: A Phylogenetic Approach. Blackwell Sciences Ltd, Oxford.

Penny, D. 1982. Towards a Basis for Classification: The Incompleteness of Distance Measures, Incompatibility Analysis and Phenetic Classification. *J. Theor. Biol.* **96**:129-142.

Posada, D., and K. Crandall. 1998. MODELTEST: Testing the Model of DNA Substitution. *Bioinformatics* **14**:817-818.

Posada, D., and K. Crandall. 2001a. Selecting the Best-Fit Model of Nucleotide Substitution. *Syst. Biol.* **50**:580-601.

Posada, D., and K. Crandall. 2001c. Selecting Models of Nucleotide Substitution: An Application to Human Immunodeficiency Virus 1 (HIV-1). *Mol. Biol. Evol.* **18**:897-906.

Rambaut, A., and M. Charleston. 2001. TreeEdit, version v1.0a8. University of Oxford.

Rambaut, A., and N. Grassly. 1997. Seq-Gen: An Application for the Monte Carlo Simulation of DNA Sequence Evolution Along Phylogenetic Trees. *Comput. Appl. Biosci.* **13**:235-238.

Rodrigo, A. 1993. Calibrating the Bootstrap Test of Monophyly. *Int. J. Parasitol.* **23**:507-14.

Rodrigo, A., P. R. Bergquist, and P. L. Bergquist. 1993. Inadequate Support for an Evolutionary Link Between the Metazoa and the Fungi. *Syst. Biol.* **43**:578-584.

Rodrigo, A., M. Kelly-Borges, P. R. Bergquist, and P. L. Bergquist. 1993. A Randomisation Test of the Null Hypothesis that two Cladograms are Sample Estimates of a Parametric Phylogenetic Tree. *N. Z. J. Bot.* **31**:257-268.

Rogers, J. 1997. On the Consistency of Maximum Likelihood Estimation of Phylogenetic Trees from Nucleotide Sequences. *Syst. Biol.* **46**:354-357.

Sanderson, M. 1995. Objections to Bootstrapping Phylogenies: A Critique. *Syst. Biol.* **44**:299-320.

Schoniger, M., and A. Von Haeseler. 1995. Performance of the Maximum Likelihood, Neighbor-Joining, and Maximum Parsimony Methods when Sequence Sites are not Independent. *Syst. Biol.* **44**:533-47.

Shibata, R. 1989. Statistical Aspects of Model Selection. Springer, New York.

Shimodaira, H., and M. Hasegawa. 1999. Multiple Comparisons of Log-Likelihood with Applications to Phylogenetic Inference. *Mol. Biol. Evol.* **16**:1114-1116.

Siddall, M. 1998. Success of parsimony in the four-taxon case: Long-branch repulsion by likelihood in the Farris Zone. *Cladistics* **14**:209-220.

Steel, M., and J. Hein. 2000. Applying the Thorne-Kishino-Felsenstein Model to Sequence Evolution on a Star-shaped Tree. *Applied Mathematic Letters* **14**:679-684.

Strimmer, K., and A. Rambaut. 2002. Inferring Confidence Sets of Possibly Misspecified Gene Trees. *Proc. R. Soc. Lond. B* **269**:137-142.

Sullivan, J., K. Holsinger, and C. Simon. 1995. Among-site Rate Variation and Phylogenetic Analysis of 12S rRNA in Sigmodontine Rodents. *Mol. Biol. Evol.* **12**:988-1001.

Sullivan, J., and D. Swofford. 1997. Are Guinea Pigs Rodents? The Inportance of Adequate Models in Molecular Phylogenetics. *Journal of Mammalian Evolution* **4**:77-86.

Sullivan, J., and D. Swofford. 2001. Should We Use Model-Based Methods for Phylogenetic Inference When We Know that Assumptions About Among-Site Rate Variation and Nulceotide Substitution Pattern are Violated. *Syst. Biol.* **50**:723-729.

Sullivan, J., D. Swofford, and G. Naylor. 1999. The Effect of Taxon Smapling on Estimating Rate Heterogeneity Parameters of Maximum Likelihood Models. *Mol. Biol. Evol.* **16**:1347-1356.

Swofford, D. 1996. PAUP* 4b10: *Phylogenetic Analysis Using Parsimony (and other methods). version 4b10. Sinauer.

Swofford, D., G. Olsen, P. Waddell, and D. Hillis. 1996. Phylogenetic Inference. Pages 407-514 *in* Molecular Systematics (D. Hillis, C. Moritz, and B. Mable, eds.). Sinauer, Sunderland, Massachusetts.

Swofford, D., P. Waddell, J. Huelsenbeck, P. Foster, P. Lewis, and J. Rogers. 2001. Bias in Phylogenetic Estimation and its Relevance to the Choice between Parsimony and Likelihood Methods. *Syst. Biol.* **51**:525-539.

Takahashi, K., and M. Nei. 2000. Efficiencies of Fast Algorithms of Phylogenetic Inference Under the Criteria of Maximum Parsimony, Minimum Evolution and Maximum Likelihood When a Large Number of Sequences are Used. *Mol. Biol. Evol.* **17**:1251-1258.

Tamura, K., and M. Nei. 1993. Estimation of the Numver of Nucleotide Substitutions in the Control Region of Mitochondrial DNA in Humans and Chimpanzees. *Mol. Biol. Evol.* **10**:512-526.

Tavare, S. 1986. Some Probabilistic and Statistical Problems on the Analysis of DNA Sequences. *Lec. Math. Life Sci.* **17**:57-86.

Templeton, A. 1983. Phylogenetic Inference from Restriction Endonuclease Cleavage Site Maps with Particular Reference to the Humans and Apes. *Evolution* **37**:221-224.

Thompson, J., T. Gibson, F. Plewniak, F. Jeanmougin, and D. Higgins. 1997. The ClustalX Windows Interface: Flexible Strategies for Multiple Sequence Alignment Aided by Quality Analysis Tools. *Nucl. Acids. Res.* **24**:4876-4882.

Waddell, P. 1995. Statistical Methods of Phylogenetic Analysis, Including Hadamard Conjugations, LogDet transforms, and Maximum Likelihood.Massey University, Palmerston North, New Zealand.

Wainright, P., G. Hinkle, M. Sogin, and S. Stickel. 1993. Monophyletic Origins of the Metazoa. *Science* **260**:340-342.

Wakeley, J. 1994. Substitution Rate Variation Among Sites and the Estimation of Transition Bias. *Mol. Biol. Evol.* **11**:436-442.

Yang, Z. 1994a. Maximum Likelihood Phylogenetic Estimation from DNA Sequences with Variable Rates over Sites: Approximate Methods. *J. Mol. Evol.* **39**:306-314.

Yang, Z. 1994b. Statistical Properties of the Maximum Likelihood Method of Phylogenetic Estimation and Comparison with Distance Matrix Methods. *Syst. Biol.* **43**:329-342.

Yang, Z. 1996a. Among-site Rate Variation and its Impact on Phylogenetic Analyses. *TRENDS in Ecology and Evolution* **11**:367-72.

Yang, Z. 1996b. Phylogenetic Analyses Using Parsimony and Likelihood Methods. *J. Mol. Evol.* **42**:294-307.

Yang, Z. 1996c. Maximum-Likelihood Models for Combined Analyses of Multiple Sequence Data. *J. Mol. Evol.* **42**:587-596.

Yang, Z. 1997. How Often do Wrong Models Produce Better Phylogenies. *Mol. Biol. Evol.* **14**:105-108.

Yang, Z., N. Goldman, and A. Friday. 1994. Comparison of Models for Nucleotide Substitution use in Maximum-Likelihood Phylogenetic Estimation. *Mol. Biol. Evol.* **11**:316-324.

Yang, Z., N. Goldman, and A. Friday. 1995. Maximum Likelihood Trees From DNA Sequences: A Peculiar Statistical Estimation Problem. *Syst. Biol.* **44**:384-399.

Zharkikh, A., and W.-H. Li. 1995. Estimation of Confidence in Phylogeny: the Complete-and-partial Bootstrap Technique. *Mol. Phylogenet. Evol.* **4**:44-63.

Zuckerkandl, E., and L. Pauling. 1962. Molecular Disease, Evolution, and Genetic Heterogeneity. Pages 189-225 *in* Horizons in Biochemistry (M. Marsha, and B. Pullman, eds.). Academic, New York.

7 Appendix

2.2.4 Estimating Trees from the Replicate Data Sets

Sample nexus files for the branch and bound searches performed to estimate the trees from the replicate data sets. The models of evolution that were used to estimate the tree from sequences are (a) JC, (b) JC+Γ, (c) HKY, (d) HKY+Γ, (e) GTR, (f) GTR+Γ.

(a)

```
begin PAUP;
log file=logJC append;
set autoclose = yes criterion = likelihood notifybeep=no errorbeep=no;
lset NST=1 basefreq=equal ;
BandB multrees=no;
savetrees brlens=yes append=yes format=nexus file=8tJC.tre;
savetrees brlens=yes append=yes format=phylip file=8tJC.phy;
gettrees mode=7 file=8tNull.nex;
lscores 1 2 / NST=1 basefreq=equal  longfmt=yes khtest=fullopt bootreps=1000 tail=1;
end;
```

(b)

```
begin PAUP;
log file=logJC+G append;
set autoclose = yes criterion = likelihood notifybeep=no errorbeep=no;
lset NST=1 basefreq=equal rates=gamma shape=estimate;
BandB multrees=no;
savetrees brlens=yes append=yes format=nexus file=8tJC+G.tre;
savetrees brlens=yes append=yes format=phylip file=8tJC+G.phy;
gettrees mode=7 file=8tNull.nex;
lscores 1 2 / NST=1 basefreq=equal rates=gamma shape=estimate longfmt=yes khtest=fullopt
bootreps=1000 tail=1;
end;
```

(c)

```
begin PAUP;
log file=logHKY append;
set autoclose = yes criterion = likelihood notifybeep=no errorbeep=no;
lset NST=2 basefreq=estimate tratio=estimate;
BandB multrees=no;
savetrees brlens=yes append=yes format=nexus file=8tHKY.tre;
savetrees brlens=yes append=yes format=phylip file=8tHKY.phy;
gettrees mode=7 file=8tNull.nex;
lscores 1 2 / NST=2 basefreq=estimate tratio=estimate longfmt=yes khtest=fullopt bootreps=1000
tail=1;
end;
```

(d)

```
begin PAUP;
log file=logHKY+G append;
set autoclose = yes criterion = likelihood notifybeep=no;
lset NST=2 basefreq=estimate tratio=estimate rates=gamma shape=estimate;
BandB multrees=no;
```

```
savetrees brlens=yes append=yes format=nexus file=8tHKY+G.tre;
savetrees brlens=yes append=yes format=phylip file=8tHKY+G.phy;
gettrees mode=7 file=8tNull.nex;
lscores 1 2 / NST=2 basefreq=estimate tratio=estimate rates=gamma shape=estimate longfmt=yes
khtest=fullopt bootreps=1000 tail=1;
end;
```

(e)

```
begin PAUP;
log file=logGTR append;
set autoclose = yes criterion = likelihood notifybeep=no errorbeep=no;
lset NST=6 basefreq=estimate Rmatrix=estimate ;
BandB multrees=no;
savetrees brlens=yes append=yes format=nexus file=8tGTR.tre;
savetrees brlens=yes append=yes format=phylip file=8tGTR.phy;
gettrees mode=7 file=8tNull.nex;
lscores 1 2 / NST=6 basefreq=estimate Rmatrix=estimate longfmt=yes khtest=fullopt bootreps=1000
tail=1;
end;
```

(f)

```
begin PAUP;
log file=logGTR+G append;
set autoclose = yes criterion = likelihood notifybeep=no errorbeep=no;
lset NST=6 basefreq=estimate Rmatrix=estimate rates=gamma shape=estimate;
BandB multrees=no;
savetrees brlens=yes append=yes format=nexus file=8tGTR+G.tre;
savetrees brlens=yes append=yes format=phylip file=8tGTR+G.phy;
gettrees mode=7 file=8tNull.nex;
lscores 1 2 / NST=6 basefreq=estimate Rmatrix=estimate rates=gamma shape=estimate longfmt=yes
khtest=fullopt bootreps=1000 tail=1;
end;
```

3.2.2 Simulating and Estimating the Replicate Data Sets

The PAUP command blocks for estimating (a) T_{ML}, (b) the likelihood of the data constrained to the true topology and (c) the likelihood of the data constrained to the true tree and the true model parameters. These command lines match the JC model of evolution and must be altered slightly to match each specific model. Each data set is executed before these commands can be performed.

(a)

```
begin PAUP;
log file = log append;
set autoclose = yes criterion = likelihood;
lset NST=1 basefreq=equal;
alltrees;
savetrees brlens=yes append = yes format = nexus file=4t.tre;
savetrees append = yes format = phylip file =4t.phy;
lscores 1 / NST=1 basefreq=equal scorefile=MLtreelscores append=yes;
end;
```

(b)

```
begin PAUP;
log file = log append;
set autoclose = yes criterion = likelihood;
```

```
lset NST=1 basefreq=equal;
loadconstr file=4Farris.nex;
alltrees enforce=yes constraints=PHYLIP_1;
savetrees brlens=yes append = yes format = nexus file=4t.tre;
savetrees append = yes format = phylip file =4t.phy;
lscores 1 / NST=1 basefreq=equal scorefile=truetoplscores append=yes;
end;
```

(c)

```
begin PAUP;
set autoclose = yes
criterion = likelihood;
gettrees storebrlens=yes file =4Farris.nex;
lscores 1 / NST=1 basefreq=equal userbrlens=yes scorefile=truetreelscores append=yes;
end;
```

3.2.5 The Analysis of Type I Error for the SOWH Test

d2alpha.pl uses the δ` values in the file DeltaStarTreeModel$_1$estModel$_2$ and calculates the Δ values from the files MLtreelscores and truetoplscores for a comparison of δ` to Δ and counts the frequency with which type I error occurs.

```
#!usr/bin/perl

open REFFILE, 'DeltaStarFarrisJCestJC';
open LNLFILE1, 'MLtreelscores';
open LNLFILE2, 'truetoplscores';
open NEWFILE, '>>Totals4alpha';
@Dstar=(<REFFILE>);
close REFFILE;

while(<LNLFILE1>){
        if (($one,$LNL) = ($_ =~ /(^1\s+)(\d+.\d+)/)){
                push (@LNL1,$LNL);
        }
}
while(<LNLFILE2>){
        if(($onemore,$LNLmore) = ($_ =~ /(^1\s+)(\d+.\d+)/)){
                push (@LNL2,$LNLmore);
        }
}
while ($counter<1000000){
        $convert1=$LNL1[$counter];
        chomp $convert1;
        $convert2=$LNL2[$counter];
        chomp $convert2;
        $difference=$convert2-$convert1;
        chomp $difference;
        push (@Delta,$difference);
        $counter+=1;
}
$replicate=0;
foreach $Dstar (@Dstar){
        &compare ($replicate);
        $comparecounter=0;
}
```

```
sub compare{
        while ($comparecounter<1000){
                $interim=$_[0];
                $conversion1=$Dstar[$interim];
                chomp $conversion1;
                $convertInterim=1000*$interim;
                chomp $convertInterim;
                $conversion2=$Delta[$comparecounter+$convertInterim];
                chomp $conversion2;
                if ($conversion1>=$conversion2){
                        $total+=1;
                        $comparecounter+=1;
                }
                else{
                        $comparecounter+=1;
                }
        }
        if ($total>=950){
                $Type1Counter++;
        }
        print NEWFILE "$total\n";
        $total=0;
        $replicate+=1;
}
print "$Type1Counter\n";
print "Operation Performed\n";
```

4.3.1 Estimating the Model of Evolution

The PAUP* commands that estimate a neighbour-joining tree and evaluate 56 nested models for use by Modeltest

```
#NEXUS
[! ***** MODELFIT BLOCK -- MODELTEST 3.0 *****]
[The following command will calculate a NJ tree using the JC69 model of evolution]

BEGIN PAUP;
        log file= modelfit.log replace;
        DSet distance=JC objective=ME base=equal rates=equal pinv=0
        subst=all negbrlen=setzero;
        NJ showtree=no breakties=random;
End;
[!
***** BEGIN TESTING 56 MODELS OF EVOLUTION ***** ]

BEGIN PAUP;
Set criterion=like;
[!** Model 1 of 56 * Calculating JC **]
lscores 1/ nst=1 base=equal rates=equal pinv=0 scorefile=model.scores replace;
[!** Model 2 of 56 * Calculating JC+I **]
lscores 1/ nst=1 base=equal rates=equal pinv=est scorefile=model.scores append;
[!** Model 3 of 56 * Calculating JC+GÊ**]
lscores 1/ nst=1 base=equal rates=gamma shape=est pinv=0 scorefile=model.scores append;
[!** Model 4 of 56 * Calculating JC+I+G **]
lscores 1/ nst=1 base=equal rates=gamma shape=est pinv=est scorefile=model.scores append;
[!** Model 5 of 56 * Calculating F81 **]
lscores 1/ nst=1 base=est rates=equal pinv=0 scorefile=model.scores append;
[!** Model 6 of 56 * Calculating F81+I **]
```

```
lscores 1/ nst=1 base=est rates=equal pinv=est scorefile=model.scores append;
[!** Model 7 of 56 * Calculating F81+G **]
lscores 1/ nst=1 base=est rates=gamma shape=est pinv=0 scorefile=model.scores append;
[!** Model 8 of 56 * Calculating F81+I+G **]
lscores 1/ nst=1 base=est rates=gamma shape=est pinv=est scorefile=model.scores append;
[!** Model 9 of 56 * Calculating K80 **]
lscores 1/ nst=2 base=equal tratio=est rates=equal pinv=0 scorefile=model.scores append;
[!** Model 10 of 56 * Calculating K80+I **]
lscores 1/ nst=2 base=equal tratio=est rates=equal pin=est scorefile=model.scores append;
[!** Model 11 of 56 * Calculating K80+G **]
lscores 1/ nst=2 base=equal tratio=est rates=gamma shape=est pinv=0 scorefile=model.scores
append;
[!** Model 12 of 56 * Calculating K80+I+G **]
lscores 1/ nst=2 base=equal tratio=est rates=gamma shape=est pinv=est scorefile=model.scores
append;
 [!** Model 13 of 56 * Calculating HKY **]
lscores 1/ nst=2 base=est tratio=est rates=equal pinv=0 scorefile=model.scores append;
[!** Model 14 of 56 * Calculating HKY+I **]
lscores 1/ nst=2 base=est tratio=est rates=equal pinv=est scorefile=model.scores append;
[!** Model 15 of 56 * Calculating HKY+G **]
lscores 1/ nst=2 base=est tratio=est rates=gamma shape=est pinv=0 scorefile=model.scores
append;
[!** Model 16 of 56 * Calculating HKY+I+G **]
lscores 1/ nst=2 base=est tratio=est rates=gamma shape=est pinv=est scorefile=model.scores
append;
[!** Model 17 of 56 * Calculating TrNef **]                    [a b c d e f]
lscores 1/ nst=6 base=equal rmat=est rclass=(a b a a e a) rates=equal pinv=0
scorefile=model.scores append;
[!** Model 18 of 56 * Calculating TrNef+I **]
lscores 1/ nst=6 base=equal rmat=est rates=equal pinv=est scorefile=model.scores append;
[!** Model 19 of 56 * Calculating TrNef+G **]
lscores 1/ nst=6 base=equal rmat=est rates=gamma shape=est pinv=0 scorefile=model.scores
append;
[!** Model 20 of 56 * Calculating TrNef+I+G **]
lscores 1/ nst=6 base=equal rmat=est rates=gamma shape=est pinv=est scorefile=model.scores
append;
[!** Model 21 of 56 * Calculating TrN **]
lscores 1/ nst=6 base=est rmat=est rates=equal pinv=0 scorefile=model.scores append;
[!** Model 22 of 56 * Calculating TrN+I **]
lscores 1/ nst=6 base=est rmat=est rates=equal pinv=est scorefile=model.scores append;
[!** Model 23 of 56 * Calculating TrN+G **]
lscores 1/ nst=6 base=est rmat=est rates=gamma shape=est pinv=0 scorefile=model.scores append;
[!** Model 24 of 56 * Calculating TrN+I+G **]
lscores 1/ nst=6 base=est rmat=est rates=gamma shape=est pinv=est scorefile=model.scores
append;
[!** Model 25 of 56 * Calculating K3P **]                      [a b c d e f]
lscores 1/ nst=6 base=equal rmat=est rclass=(a b c c b a) rates=equal pinv=0
scorefile=model.scores append;
[!** Model 26 of 56 * Calculating K3P+I **]
lscores 1/ nst=6 base=equal rmat=est rates=equal pinv=est scorefile=model.scores append;
[!** Model 27 of 56 * Calculating K3P+G **]
lscores 1/ nst=6 base=equal rmat=est rates=gamma shape=est pinv=0 scorefile=model.scores
append;
[!** Model 28 of 56 * Calculating K3P+I+G **]
lscores 1/ nst=6 base=equal rmat=est rates=gamma shape=est pinv=est scorefile=model.scores
append;
[!** Model 29 of 56 * Calculating K3Puf **]
lscores 1/ nst=6 base=est rmat=est rates=equal pinv=0 scorefile=model.scores append;
[!** Model 30 of 56 * Calculating K3Puf+I **]
lscores 1/ nst=6 base=est rmat=est rates=equal pinv=est scorefile=model.scores append;
```

```
[!** Model 31 of 56 * Calculating K3Puf+G **]
lscores 1/ nst=6 base=est rmat=est rates=gamma shape=est pinv=0 scorefile=model.scores append;
[!** Model 32 of 56 * Calculating K3Puf+I+G **]
lscores 1/ nst=6 base=est rmat=est rates=gamma shape=est pinv=est scorefile=model.scores
append;
[!** Model 33 of 56 * Calculating TIMef **]                    [a b c d e f]
lscores 1/ nst=6 base=equal rmat=est rclass=(a b c c e a) rates=equal pinv=0
scorefile=model.scores append;
[!** Model 34 of 56 * Calculating TIMef+I **]
lscores 1/ nst=6 base=equal rmat=est rates=equal pinv=est scorefile=model.scores append;
[!** Model 35 of 56 * Calculating TIMef+G **]
lscores 1/ nst=6 base=equal rmat=est rates=gamma shape=est pinv=0 scorefile=model.scores
append;
[!** Model 36 of 56 * Calculating TIMef+I+G **]
lscores 1/ nst=6 base=equal rmat=est rates=gamma shape=est pinv=est scorefile=model.scores
append;
[!** Model 37 of 56 * Calculating TIM **]
lscores 1/ nst=6 base=est rmat=est rates=equal pinv=0 scorefile=model.scores append;
[!** Model 38 of 56 * Calculating TIM+I **]
lscores 1/ nst=6 base=est rmat=est rates=equal pinv=est scorefile=model.scores append;
[!** Model 39 of 56 * Calculating TIM+G **]
lscores 1/ nst=6 base=est rmat=est rates=gamma shape=est pinv=0 scorefile=model.scores append;
[!** Model 40 of 56 * Calculating TIM+I+G **]
lscores 1/ nst=6 base=est rmat=est rates=gamma shape=est pinv=est scorefile=model.scores
append;
[!** Model 41 of 56 * Calculating TVMef **]                    [a b c d e f]
lscores 1/ nst=6 base=equal rmat=est rclass=(a b c d b e) rates=equal pinv=0
scorefile=model.scores append;
[!** Model 42 of 56 * Calculating TVMef+I **]
lscores 1/ nst=6 base=equal rmat=est rates=equal pinv=est scorefile=model.scores append;
[!** Model 43 of 56 * Calculating TVMef+G **]
lscores 1/ nst=6 base=equal rmat=est rates=gamma shape=est pinv=0 scorefile=model.scores
append;
[!** Model 44 of 56 * Calculating TVMef+I+G **]
lscores 1/ nst=6 base=equal rmat=est rates=gamma shape=est pinv=est scorefile=model.scores
append;
[!** Model 45 of 56 * Calculating TVM **]
lscores 1/ nst=6 base=est rmat=est rates=equal pinv=0 scorefile=model.scores append;
[!** Model 46 of 56 * Calculating TVM+I **]
lscores 1/ nst=6 base=est rmat=est rates=equal pinv=est scorefile=model.scores append;
[!** Model 47 of 56 * Calculating TVM+G **]
lscores 1/ nst=6 base=est rmat=est rates=gamma shape=est pinv=0 scorefile=model.scores append;
[!** Model 48 of 56 * Calculating TVM+I+G **]
lscores 1/ nst=6 base=est rmat=est rates=gamma shape=est pinv=est scorefile=model.scores
append;
[!** Model 49 of 56 * Calculating SYM **]                      [a b c d e f]
lscores 1/ nst=6 base=equal rmat=est rclass= (a b c d e f) rates=equal pinv=0
scorefile=model.scores append;
[!** Model 50 of 56 * Calculating SYM+I **]
lscores 1/ nst=6 base=equal rmat=est rates=equal pinv=est scorefile=model.scores append;
[!** Model 51 of 56 * Calculating SYM+G **]
lscores 1/ nst=6 base=equal rmat=est rates=gamma shape=est pinv=0 scorefile=model.scores
append;
[!** Model 52 of 56 * Calculating SYM+I+G **]
lscores 1/ nst=6 base=equal rmat=est rates=gamma shape=est pinv=est scorefile=model.scores
append;
[!** Model 53 of 56 * Calculating GTR **]
lscores 1/ nst=6 base=est rmat=est rates=equal pinv=0 scorefile=model.scores append;
[!** Model 54 of 56 * Calculating GTR+IÊ**]
lscores 1/ nst=6 base=est rmat=est rates=equal pinv=est scorefile=model.scores append;
```

```
[!** Model 55 of 56 * Calculating GTR+G **]
lscores 1/ nst=6 base=est rmat=est rates=gamma shape=est pinv=0 scorefile=model.scores append;
[!** Model 56 of 56 * Calculating GTR+I+G **]
lscores 1/ nst=6 base=est rmat=est rates=gamma shape=est pinv=est scorefile=model.scores
append;
LOG STOP;
END;
[!*** END OF MODELTEST BLOCK ***]
```

The PAUP* commands for estimating the tree and parameters using the model of evolution selected by Modeltest for the Bird *mt*DNA data set.

```
BEGIN PAUP;
log file=log;
set autoclose=yes criterion=likelihood notifybeep=no;
Lset Base=estimate Nst=6 Rmat=estimate Rates=gamma shape=estimate Pinvar=estimate;
alltrees;
savetrees brlens=yes append=yes format=phylip file=Bird.phy;
Lscores 1/ scorefile=Bird.sp append=yes;
END;
```