MASSEY UNIVERSITY
GRADUATE RESEARCH SCHOOL

Declaration Confirming Content of Digital Version of Thesis

I confirm that the content of the digital version of this thesis

Title: Lineage Specific Evolution and Phylogenetic Analysis

is the final amended version following the examination process and is identical to this hard bound paper copy.

Student's Name: Liat Shavit Grievink

Student's Signature: Liat Shavit Grievink

Date: 12/8/09

LINEAGE SPECIFIC EVOLUTION AND PHYLOGENETIC ANALYSIS

A thesis presented in partial fulfillment of the requirements for the degree

of

Doctor of Philosophy

in

Biomathematics

at Massey University, Palmerston North,

New Zealand.

Liat Shavit Grievink

2009

ABSTRACT

Phylogenetic models generally assume a homogeneous, time reversible, stationary process. These assumptions are often violated by the real, far more complex, evolutionary process. This thesis is centered on non-homogeneous, lineage-specific, properties of molecular sequences. It consist several related but independent studies. LineageSpecificSeqgen, an extension to the Seq-Gen program, which allows generation of sequences with changes in the proportion of variable sites, is introduced. This program is then used in a simulation study showing that changes in the proportion of variable sites can hinder tree estimation accuracy, and that tree reconstruction under the best-fit model chosen using a relative test can result in a wrong tree. In this case, the less commonly used absolute model-fit was a better predictor of tree estimation accuracy. This study found that increased taxon sampling of lineages that have undergone a change in the proportion of variable sites was critical for accurate tree reconstruction and that, in contrast to some earlier findings, the accuracy of maximum parsimony is adversely affected by such changes.

This thesis also addresses the well-known long-branch attraction artifact. A non-parametric bootstrap test to identify changes in the substitution process is introduced, validated, and applied to the case of Microsporidia, a highly reduced intracellular parasite. Microsporidia was first thought to be an early branching eukaryote, but is now believed to be sister to, or included within, fungi. Its apparent basal eukaryote position is considered a result of long-branch attraction due to an elevated evolutionary rate in the microsporidian lineage. This study shows that long-branch estimates and basal positioning of Microsporidia both correlate with increased proportions of radical substitutions in the microsporidian lineage. In simulated data, such increased proportions of radical substitutions leads to erroneous long-branch estimates. These results suggest that the long microsporidian branch is likely to be a result of an increased proportion of radical substitutions on that branch, rather than increased evolutionary rate *per se*.

The focus of the last study is the intriguing case of Mesostigma, a fresh water green alga for which contradicting phylogenetic relationships were inferred. While some studies placed Mesostigma within the Streptophyta lineage (which includes land plants), others placed it as the deepest green algae divergence. This basal positioning is regarded as a result of long-branch attraction due to poor taxon sampling. Reinvestigation of a 13-taxon mitochondrial amino acid dataset and a sub-dataset of 8 taxa reveals that site sampling, and in particular the treatment of missing data, is just as important a factor for accurate tree reconstruction as taxon sampling. This study identifies a difficulty in recreating the long-branch attraction observed for the 8-taxon dataset in simulated data. The cause is likely to be the smaller number of amino acid characters per site in simulated data compared to real data, highlighting the fact that there are properties of the evolutionary process that are yet to be accurately modeled.

ACKNOWLEDGMENT

First and foremost, I would like to thank my supervisors Dr. Barbra Holland, Prof. David Penny, and Prof. Mike Hendy for allowing me to join their research group, for their expertise, guidance, suggestions, and encouragement. I am grateful for the opportunities they have offered me and helped me realize. This work would not have been possible without their open-door policy and immense patience. I would also like to acknowledge the financial support from the Marsden fund given to Barbara.

I am extremely grateful to Prof. Pete Lockhart for his interest in my study, extensive discussions, many valuable ideas and suggestions, and for his inspiring enthusiasm.

My sincere thanks go to Prof. Bill Martin for his involvement in this work, for sharing his ideas with me, for his kind hospitality, and for the financial support that has enabled my two visits to Dusseldorf.

Thanks also go to Dr. Tal Dagan, who has kindly shared her office in Dusseldorf with me, and other members of Bill's lab for helping me find my way around Dusseldorf.

I am grateful to Prof. David Bryant for his contribution to this study, and his great efforts to explain things clearly and simply.

Special thanks are due to Dr. Klaus Schliep for his keen help with R and Latex, Warwick Allen for getting me started with Perl, and Tim White for programming advice. I am also thankful to my office-mates Angela, Atheer, and Bennet, and other colleagues, for their company, discussions, and coffee breaks. Thanks also go to Susan Adams, Joy Wood, and Karen Sinclair for their kind help and support.

Constructive comments and suggestions from many participants of the NZ phylogenetic meeting, in the past 3 years, were very much appreciated. In particular, I would like to thank Prof. Mike Steel for helpful discussions.

I am most grateful to my friends, in Israel, The Netherlands, and here in New Zealand, who supported me throughout this study, and who accepted my limited social (and e-mailing) time slots. Special gratitude goes to Ofir, Evelyn, Nell, Aurelie, Estelle, and the lunch-time gang for the emotional support and their very much valued friendship.

Last, but not least, I would like to thank my close and extended family. I am greatly indebted to my husband, Hilbert. He has kept me grounded and sane through this journey, helped me keep things in perspective, celebrated my small successes with me, stood by me, and provided me with love, support, and encouragement, all while undertaking his own PhD study. I thank my parents, Rachel and Meir, for their love and support throughout my life and for their understanding when I decided to study on the other side of the world. Their belief in my ability to do anything I put my mind to gave me the determination to complete this study. This thesis is humbly dedicated to them.

CONTENTS

Chapter 1

Introduction

1.1 Candidate's Note

This thesis is a collection of research papers, either published, accepted for publication, or in preparation for submission. Each of chapters 2-5 is a self-contained paper and can be read in a stand-alone manner; the thesis therefore contains some repetition and differences in format. I have, however, standardized the format as much as possible and added internal referencing where appropriate. No changes were made in the content of published papers. The paper bound at the back of this thesis as an appendix is a result of a study which I started as part of my Masters thesis. However, I spent much of the first year of my PhD study extending and preparing this work for publication. In particular, the work involving corrected-maximum parsimony (including its implementation) was done entirely during my PhD study. This paper is thus included here and should be included, in part, in the assessment of my PhD research. This thesis is a report on the progress made during the three years of my PhD study. Some of the projects described are still on-going, and of course further research will stem from this work.

Much of the work presented in this thesis is a result of collaborative projects. Nevertheless, this work is my own. I have done the vast majority of the work for each of the papers, including all the programming and analyses. I also had the responsibility for writing each manuscript. All the papers included here greatly benefited from invaluable discussions with my supervisors: Dr. Barbara Holland, Prof. David Penny and Prof. Mike Hendy, and discussions with Prof. Pete Lockhart. Discussions with my collaborators Prof. David Bryant, Prof. Bill Martin and Prof. Pete Lockhart were extremely valuable for the work presented in Chapter 4. In particular, the initial idea and formulation of the bootstrap test is of Prof. David Bryant.

1.2 **Overview**

From the time of Charles Darwin, biologists have sought to reconstruct the evolutionary relationships of all life on Earth (both living and extinct) and express it in the form of a phylogenetic tree. This requires reliance on mathematical models to describe the evolutionary process. A major problem biologists are faced with is that reality is too complex for the math to handle. Almost all mathematical models of evolution assume that the same processes act over all parts of the tree. They fail to account for the fact that sequences in different lineages acquire their own particular properties. This thesis is concerned with the effect this over-simplification has on the estimation of evolutionary relationships and our understanding of the process of evolution. The aim of this introductory chapter is to give a brief overview of the motivation for the work, and to describe the progress that is presented in this thesis.

1.3 Basic concepts

1.3.1 Phylogenetic trees

A phylogenetic tree is a graph composed of nodes and branches, in which any two nodes are connected by a unique path. Nodes on the tree represent taxonomic units (species, populations, individuals), and branches define the relationship between taxa in terms of descent. Internal nodes on the tree correspond to speciation events. In general, the process of speciation is assumed to be binary, so that each speciation results in two new species and the tree is bifurcating. External nodes, also called terminal nodes, leaves, or tips, usually correspond to extant (living) taxonomic units. In molecular phylogeny, branch lengths typically represent the number of changes in the molecular sequence (DNA, RNA, or amino acids) that have occurred on that branch. The branching pattern of the tree is called a topology. A phylogenetic tree can either be rooted or unrooted. Rooted trees have a node, called the root, which represents the common ancestor from which a unique directed path leads to any other node. An unrooted tree characterizes the relationships between taxa; but an evolutionary path and a common ancestor are not defined. Figure 1.1 illustrates these concepts. A common notation to represent a phylogenetic tree is the Newick format (see http://evolution.genetics.washington.edu/phylip/newicktree.html). Using this format, the tree in Figure 1.1b can be written as (((A,B),C),D),(E,F)). The work presented in this thesis is limited to trees and therefore does not consider lateral gene transfer (transfer of genetic material between different species [1]), recombination, or hybridization, which cannot be represented by a tree.

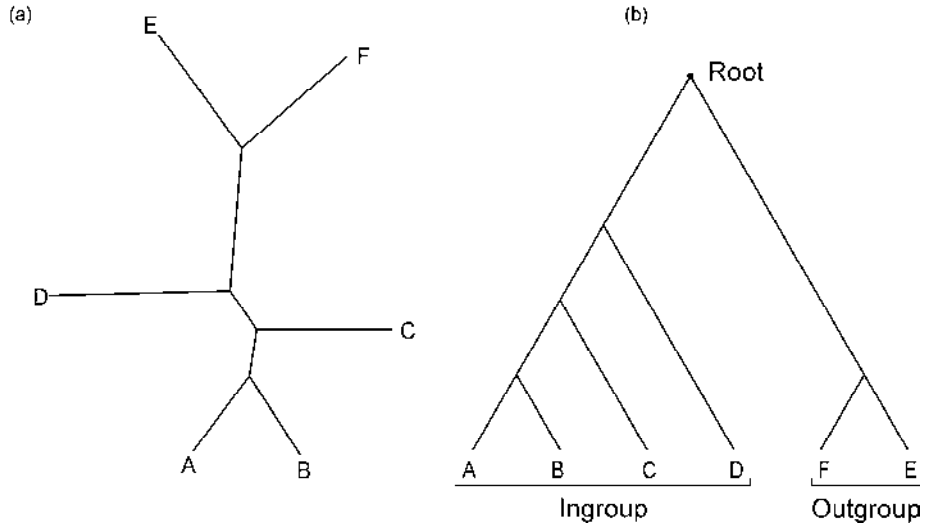


Figure 1.1 – Examples of unrooted (a) and rooted (b) 6-taxon binary trees. The taxa set {A, B, C, D} is the ingroup, the outgroup taxa {E, F} can be used to root the tree (b).

An unrooted tree can be rooted in two ways. The first is using an additional taxon (or group of taxa) called an outgroup; the taxa that do not belong to the outgroup are called the ingroup, and the internal node where the outgroup joins the ingroup is designated as the root of the ingroup tree. A second way of rooting an unrooted tree is using the molecular clock assumption. In this case, the rate of change is assumed to be constant and branch lengths therefore correspond to elapsed time. The point on the tree from which the distance (branch lengths) to all tips is approximately equal is then named the root of the tree [2]. The molecular clock assumption can be relaxed, allowing rates to vary across the tree [3, 4, 5].

Evolution takes place when the allele frequencies within a population change over time. Such changes occur due to mutations and their fixation in the population. Mutations can arise as a result of unrepaired copying errors (in the DNA) during cell division or as a result of exposure to radiation, chemicals, or viruses. While some mutations become fixed (wide-spread) in the population, others do not; this may be random (if the changes are neutral) or may be dependent on functional and structural constraints. A common mutation is a single nucleotide replacement. Mutations of this type that have reached fixation are known as substitutions. In molecular phylogenetics trees are usually reconstructed using the observed substitutions in the molecular sequence.

1.3.2 Sequence alignments

Sequence alignments are the typical input data for phylogenetic tree estimation methods. A sequence alignment is an arrangement of DNA, RNA, or amino acids sequences where each sequence is represented as a row in a matrix. Gaps can be inserted between characters in a sequence so that assumed homologous sites (those which have evolved from a common ancestor) are aligned in columns. A site that has the same character state for all the aligned sequences is called 'constant' or 'invariant'. Alignments are often done by computer programs which are based on dynamic programming algorithms, employing an explicit optimization function that rewards matches and penalizes mismatches, insertions, and deletions [6, 7]. As sequence alignments are at the base of all phylogenetic analysis, their correctness is critical [8]. One objective way of assessing the quality of an alignment is the heads-or-tails technique [9].

1.3.3 Tree reconstruction methods

Existing phylogenetic methods can be classified into three categories: parsimony, distance-based, and likelihood-based methods. These methods consist of two components: an optimality criterion, and a search strategy. The optimality criterion assigns a score to a given tree and data. For parsimony methods the score is the minimum number of substitutions required to explain the data assuming the sequences evolved on the given tree. In the case of distance methods, two optimality criteria are commonly used: minimum evolution (where the score is the length of the tree), and least squares (where the score is the some of square differences between expected and observed distances). However, in some distance methods (those that use clustering algorithms) the optimality criterion is combined with the search strategy (see below). For likelihood-based methods, such as maximum likelihood or Bayesian analysis, the score is the probability of observing the data assuming the sequences evolved on the given tree according to some model of substitution, or vice versa, respectively. Unlike other methods (such as maximum parsimony or maximum likelihood), Bayesian analysis does not attempt to find a single best-tree; it generates an approximation of the

posterior probability distribution of all parameters (i.e. tree shape, branch lengths, and model) typically using Markov Chain Monte Carlo (for more information see [10, 11]).

The search strategy is an algorithm for searching through the space of all trees. Exhaustive search is possible when the tree space is small (small number of taxa), however it becomes difficult as the number of taxa increases [7]; this is because the number of possible trees increases super-exponentially in relation to the number of taxa [2]. Therefore a heuristic search is normally used, where an initial tree is constructed (typically by a distance method) and is improved upon until a local optimal tree is reached. This tree is not guaranteed to be the global optimal tree [12].

In some distance methods (e.g. neighbor joining) the optimality criterion and the search strategy are combined, and the tree is usually constructed using a greedy algorithm. At any stage a criterion determines which two taxonomic units should be grouped together. The distance matrix is then updated (where the grouped units are represented as a single unit), and the process is repeated until only one taxonomic unit exists. See [13], for a review, or [2] for more detail on phylogenetic methods.

1.3.4 Confidence assessment

Generally (with simulations being the exception), a phylogenetic tree is an estimate of an unknown phylogeny. An important question is how good this estimate is. Confidence assessment in phylogenetics is traditionally done using bootstrap analysis. Bootstrap is a statistical technique for empirically estimating the variability of an estimate [14, 15]. The distribution of variability of the estimate is approximated by sampling with replacement from the original dataset, creating multiple datasets of the same size and inferring the phylogeny from each. The sample of phylogenies should display roughly the same variation as a sample obtained by collecting the same amount of new sites [2]. The support for each branch in the original tree is calculated as the frequency with which it is observed in the replicate trees. If a branch has strong support, it will be supported by at least some positions in each of the bootstrap samples, and all the bootstrap samples will yield this branch (for more information see [16]). While

bootstrap analysis can detect variability in the estimate due to sampling error (lack of data), it is dependent on the method and model used and is thus prone to systematic biases; if the method used is biased, the bootstrap support values will be biased too. This weakness of the use of bootstrap for confidence assessment has been highlighted by Phillips et al. [17] who showed that 100% bootstrap support can be obtained for each of two different tree topologies using the same dataset under different substitution models. In a Bayesian framework, the posterior probabilities can be used as support measures [10]. However, as in the case of bootstrap, these support values are also conditional on the model assumptions [18].

1.3.5 Substitution models

In the context of molecular phylogenetics, statistical models are used to make predictions about the substitution process along the branches of the tree. A substitution model describes in probabilistic terms the process by which one sequence of characters (nucleotides or amino acids) changes into another over time, as well as expected character state frequencies. These models are utilized by likelihood-based methods to evaluate the probability of the tree and branch lengths given the data and vice versa, and by other (generally distance-based, but sometimes parsimony) methods where they are used to correct for undetectable multiple substitutions at a site.

Phylogenetic models typically assume a homogeneous, stationary and time reversible Markovian process (the future state at a site depends solely on the current state, not on previous states). In the context of phylogenetics, the homogeneity assumption implies that the instantaneous rate matrix is constant over an edge (local homogeneity) or over the entire tree (global homogeneity). The stationarity assumption implies that the marginal probabilities of the characters (nucleotides, or amino acids) remain constant over all nodes of the tree. Finally, the reversibility assumption implies that the rate of substitution from character *i* to character *j* is the same as the rate of substitution from character *j* to character *i*.

Most nucleotide substitution models belong to the general time-reversible (GTR) [19] family. For these models there are six possible substitution types for the four nucleotides (A, C, G, T). The instantaneous rate matrix for the GTR model is: $Q = \begin{pmatrix} - & r_{AC}\Pi_C & r_{AG}\Pi_G & r_{AT}\Pi_T \\ r_{AC}\Pi_A & - & r_{CG}\Pi_G & r_{CT}\Pi_T \\ r_{AG}\Pi_A & r_{CG}\Pi_C & - & r_{GT}\Pi_T \\ r_{AT}\Pi_A & r_{CT}\Pi_C & r_{GT}\Pi_G & - \end{pmatrix}$, where the Π_i values are the equilibrium frequencies of the four nucleotides, $r_{ij} = r_{ji}$ is the rate of substitution between nucleotides *i* and *j* (this equality is required under the time reversibility assumption), and the diagonals are chosen so that the sum of values in each row is 0. Other models belonging to the GTR family are special cases of the GTR model with some constraints on its rates and frequencies parameters. For a review of models for nucleotide evolution see [20].

Unlike nucleotides, amino acids substitution models are generally based on empirical data. Such empirical matrices include for example the JTT [21], WAG [22], and Dayhoff [23] matrices. Other empirical matrices exist, with some being specific for mitochondria or chloroplast data. Amino acids substitution matrices usually consist of 190 relative rates of substitution (reversibility is assumed, and diagonal elements are fixed so that each row sums to zero).

In addition to the substitution rates and base frequencies, most modern phylogenetic reconstruction programs incorporate an option to account for variation in rates across sites. This is done using the rates-across-sites (RAS) model, sometimes specifically referred to as the 'gamma model', which was first introduced for use in phylogenetics by Yang [24]. This model allows sites to evolve at different rates by assigning rates to sites according to a gamma distribution (other distributions are possible, see [25]). Each site remains in the same rate class across the entire tree. For ease of computation, the mean rate of all sites is assumed to be 1. The number of sites with the various rates of substitutions determines the shape of the distribution which is summarized by the shape parameter (α). If the rate variation between sites is large, the shape parameter is expected to be small ($\alpha<1$), whereas if most sites evolve with similar substitution rates

the shape parameter is expected to be large (α>20). The special case of equal substitution rates across sites implies an infinite shape parameter.

Sometimes, a proportion of invariable sites (I) is also used to model some rate heterogeneity across sites, accounting for sites that are conserved (invariable) in all the homologous sequences in the dataset. The covarion model, which was first described by Fitch and Markowitz [26], allows for a changing positions of variable sites through time; where invariable sites may become variable and vice versa. Tuffley and Steel [27] introduced a hidden Markov implementation of this model. However, unlike the original description of the covarion model, the heterogeneity between lineages in this implementation is limited as it assumes a fixed proportion of invariable sites (Figure 1.2). This model has been extended to incorporate variation in rates-across-sites [28, 29, 30].

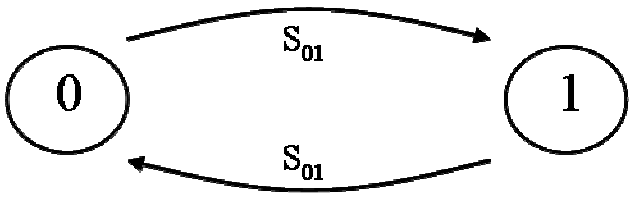


Figure 1.2 – An illustration of the covarion model of Tuffley and Steel [27]. A fixed proportion $\sigma = \frac{S_{01}}{S_{01} + S_{10}}$ **of sites are invariable, but invariable sites may become variable (with probability** S_{01}**) and vice versa (with probability** S_{10}**).**

1.3.6 The nature of the evolutionary process

The evolutionary process, as it occurs in nature, is one where selective constraints vary over time and along the sequence. Such changes in constraints can occur, for example, when a protein obtains a new function [31] or as a result of alterations in protein-protein interactions [32, 33, 34, 35]. These constraints determine, at any given time, what (proportion and positions of) sites are free to vary and the types of substitutions that can occur at any site. Unlike the processes described by standard phylogenetic models, the process of evolution is non-homogeneous. Although these biological properties have

been known since 1970 [26, 36] and seem to be a prevalent feature of molecular data [26, 35, 37, 38], existing phylogenetic models do not allow for lineage-specific proportions of variable sites and changes in the types of substitutions that can occur at each site at any given time. Such lineage-specific processes can cause what is sometimes called “heterotachy” [35], where for a specific site lineage-specific rates of substitutions are inferred.

1.3.7 Model misspecification

Models of character evolution are at the base of all phylogenetic analysis. Even methods that do not appear to use an explicit model (such as parsimony methods) make assumptions about the evolutionary process [2, 39]. Generally, models are a simplification of the true processes. In order to be useful, models should closely approximate the unknown reality, rather than describe it exactly. A more exact description of a process can always be achieved through the use of additional parameters. However, overparameterization should be avoided, as it may result in poor estimation of these parameters reducing the power of the model and its usefulness for making predictions about additional data [40]. Overparameterization may also lead to non-identifiability; a case where a tree and the parameters of a model cannot be determined from the expected distribution of the data (for instance, Matsen and Steel [41] discuss an example of a mixture model of two trees, with the same topology but different branch lengths, that produces identical expected site pattern frequencies as a third tree which has a different topology). On the other hand, model misspecification (when a model is poor approximation of reality) may systematically bias the analysis which can result in inaccurate (but sometimes apparently well-supported) estimations [42].

As model adequacy is important for correct tree estimation, several methods to evaluate model’s fit to the data have been developed. These can be divided into two categories: model selection methods and model adequacy assessment methods. Model selection methods (such as the likelihood ratio test, Akaike information criterion [43], and Bayesian information criterion [44]) choose the relative best-fit model, a model that

maximizes the likelihood of the data given the tree considering (and in most cases penalizing for) the number of parameters, from a given set of models. It is important to note that the relative best-fit model is not necessarily adequate for tree reconstruction. Model adequacy assessment methods (such as those described in [45, 46]) evaluate how well a certain model performs in predicting future observations. This is usually done by simulating predictive observations under the model in question, and comparing these to the original data using some test statistic. Unlike model selection methods, these evaluate the absolute adequacy of the model and can reject the best-fit model if some component of the evolutionary process is not accounted for in the set of models tested [47]. Until recently (2009), an implementation for model adequacy assessment methods was not available, and researchers wanting to apply these methods needed to write their own code. As a result, these methods are not yet in common use in phylogenetic analyses. An implementation of a model adequacy assessment method in a Bayesian environment is now available [48].

1.3.8 Long-branch attraction

Long-branch attraction [49, 50, 51] is a common systematic error where two non-adjacent long branches are mistakenly grouped together. This artifact is of particular concern when a distant outgroup is used in tree reconstruction for rooting and molecular dating. In this case, the long-branch lineages of the ingroup are ‘attracted’ by the outgroup lineage. This can cause an artificial early emergence of the long-branch lineages of the ingroup [52, 53]. Long-branch attraction has been suggested to affect tree reconstruction of many groups, including early Eukaryotes [54] (discussed in Chapter 4) and angiosperms [55] (discussed in the appendix). Felsenstein [49] showed that unequal rates of substitution can cause long-branch attraction and mislead tree-building methods based on parsimony. Nevertheless, Hendy and Penny [50] have shown that methods can be misled even under the molecular clock assumption (i.e. equal rates). They suggested that it is not the unequal rates *per se* that cause methods to converge to the wrong tree, but rather the estimated numbers of substitutions along the edges (similar number of substitution in non sister lineages and different number of substitutions in sister lineages). This estimation is model-dependent; therefore, long-

branch attraction is primarily caused by model misspecification (whether the model assumptions are explicit or implicit). Although Hendy and Penny [50] have shown that unequal rates are not a pre-requisite for long-branch attraction, Felsenstein's first interpretation of long branches, which was captivatingly simplistic, has stuck and long branches are often described as fast evolving lineages (e.g. [56, 57]).

1.4 A missing piece

Like several other assumptions (such as site independence, which is not dealt with in this thesis), the homogeneity assumption which is incorporated in the vast majority of character substitution models is known to be inaccurate. The evolutionary process is much more complex than that captured by the models. In particular, the proportion of variable sites is known to evolve in a lineage-specific manner (see Figure 1.3) due to changes in functional and structural constraints [35, 38]. This violates one of the main assumptions of the covarion implementation of Tuffley and Steel [27], but is closer to the original covarion idea of Fitch and Markowitz [26]. While other authors have studied non-homogeneous models (for example in the context of identifying selection [58, 59]), none of these studies allowed non-homogeneous proportion of variable sites which is the extension considered in this thesis. Variations in evolutionary constraints are also expected to cause changes in the substitution process (see Figure 1.4) leading to lineage-specific relative rates of substitutions. These features of the evolutionary process are not accounted for by current phylogenetic models. This can lead to over- or under-estimation of the number of substitutions along a branch and may cause long-branch attraction artifact. The extent to which this model misspecification affects tree reconstruction is still unknown. This thesis is focused around these lineage-specific properties, how they can be simulated, what effect they have on tree reconstruction, and perhaps most importantly can they be detected in real datasets?

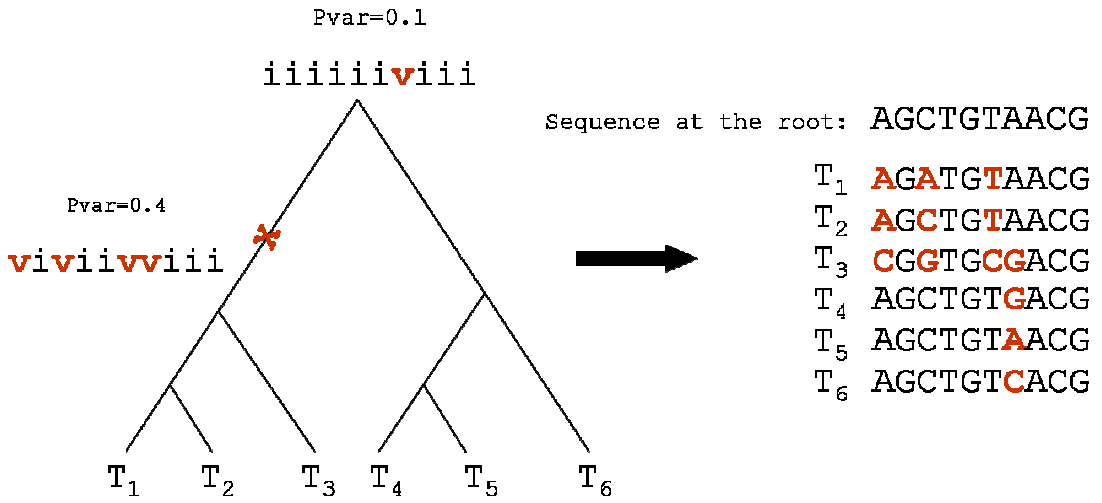


Figure 1.3 – Illustration of lineage specific proportion of variable sites. 'i' = invariable site, 'v' = variable site. At the root, the proportion of variable sites (Pvar) is 0.1. A change in Pvar (marked as X on the tree) occurs on the lineage leading to taxa T1, T2, and T3 where Pvar is 0.4. The resulting Pvar in the 6-taxon sequence alignment is then 0.4. However, this Pvar does not accurately describe the evolutionary process for taxon T4, T5, and T6 where Pvar=0.1.

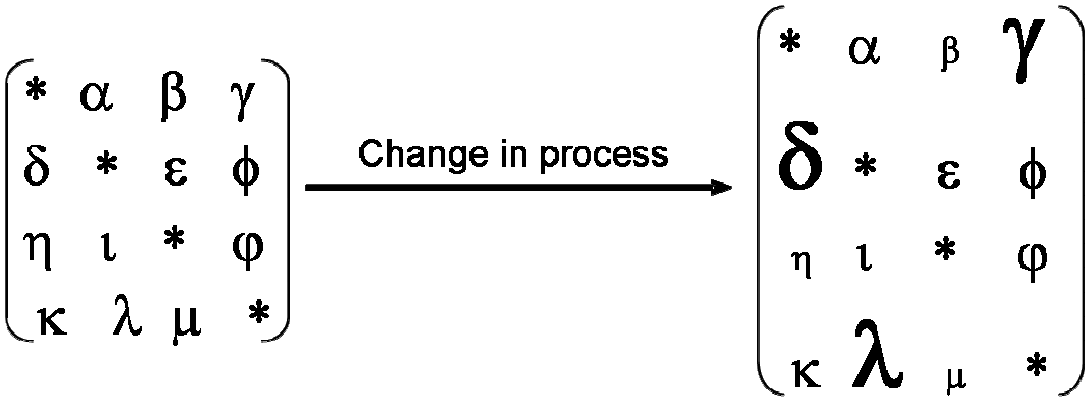


Figure 1.4 - Change in the substitution process. The matrices represent the instantaneous rate matrix. When a change in the substitution process occurs (due to changes in constraints), some types of substitutions become less frequent, some remain the same, and others become more frequent. In this diagram the type size changes illustrate changes in the value of the corresponding parameters.

1.5 Thesis Outline

Commonly used phylogenetic sequence generators employ homogeneous models of molecular sequence evolution, ignoring lineage-specific proportion of variable sites. In **Chapter 2**, I describe a new simulation tool called "LineageSpecificSeqgen" that allows systematic generation of sequences with changes in the proportion of variable sites through time. It extends the standard covarion model [27] which assumes a constant proportion of variable sites. This facilitates a more biochemically realistic simulation of the evolutionary process. Simulated sequences are used in many applications, including hypothesis testing [46], performance comparison of different tree estimation methods under various models and parameters [53, 60, 61], testing model misspecification effects on tree reconstruction [62, 63, 64], development of new models and methods [65, 66], and approximate Bayesian inference [67]. The value of these is greatly dependent on the ability of the simulation to generate data in ways that closely model the underlying biological processes. The simulator described enables testing of current models of evolution on sequences that have undergone lineage-specific evolution, as well as development of new methods to identify such processes in real data and means to account for such processes. This chapter has been published in BMC Evol. Biol. [68].

Chapter 3 includes a simulation study which explores the effect of lineage-specific proportions of variable sites on model-fit and tree-estimation accuracy. Using the LineageSpecificSeqgen simulator described in Chapter 2, this study compares tree reconstruction accuracies of five current models of nucleotide sequence evolution in a Bayesian framework, as well as the accuracy of maximum parsimony. These are applied to data containing increasing levels of change in the proportions of variable sites with and without changing positions of variable sites. Such changes can lead to the inference of lineage-specific rates of substitution at a site (heterotachy). In a *Nature* paper published in 2004 [63] (see also [69]) Kolaczkowski and Thornton claimed that maximum parsimony outperforms maximum likelihood and Bayesian analysis, and declared that maximum parsimony is unaffected by heterotachy. However, their conclusion was based on one specific case which is considered unrealistic [70]. Later

work [71, 72, 73] provided some contradictory evidence. The study presented in this chapter establishes that maximum parsimony is adversely affected by a changing proportion of variable sites, a biochemically realistic process that leads to heterotachy. This type of heterotachy was also found to hinder tree reconstruction estimation in two of the five models tested in a Bayesian framework. This study also demonstrates the importance of absolute, as opposed to relative, model adequacy assessment. Interestingly, the model with the best relative fit was sometimes found to perform worst in tree reconstruction. Absolute goodness-of-fit was found to be a good prediction tool for tree estimation accuracy. This chapter has been accepted for publication in Syst. Biol. (April, 2009) pending some revisions.

The study presented in **Chapter 4** deals with long-branch attraction and its possible causes. This study shows that long-branches, rather than being an indication of fast rates of substitution, can be a result of relaxation in evolutionary constraints manifested in a higher proportion of radical amino acids substitutions (that is, substitutions between amino acids with different chemical properties; see chapter). I present a novel test that can be used as a tool to identify variations in the substitution process in sequence data. This test is then applied to the case of Microsporidia whose tendency to branch deep in phylogenetic analysis has been tied to long-branch attraction artifacts. I intend to extend this work as part of my post-doctoral research.

Chapter 5 is centered on the robustness of phylogenetic methods to model misspecification, taxon sampling, site sampling, and missing data in the sequence alignment. I found that taxon sampling alone cannot explain the early emergence of *Mesostigma* (a species of fresh water green algae) as a sister lineage to all other green algae and that missing data in the sequence alignment significantly affects the estimated phylogeny. This study identifies a gap between simulated data, based on relative best-fit model, and real data. Particular incongruence is noted in the number of different amino acids characters per site (smaller averaged number for real data compared to simulated data). This is, in part, the result of underestimation of the proportion of invariable sites. Evolutionary constraints on the possible types of substitutions, which are unaccounted for in the common substitution models, can explain these findings.

The paper bound as an **appendix** describes a simulation study which focuses on tree estimation for rapid radiations. Such cases, where there is a combination of short internal and long external branches, are known to be prone to long-branch attraction artifacts. The performance and accuracy of several phylogenetic methods are evaluated. Biases towards specific tree topologies were identified in Maximum-likelihood, corrected- and uncorrected-neighbor-joining and corrected- and uncorrected-parsimony. This study shows that tree estimation using a single-taxon outgroup often disrupts an otherwise correct ingroup topology. Tree estimation using a two-taxon outgroup was more accurate than when using a single-taxon outgroup. However, the ingroup was most accurately recovered when no outgroup was used. This work has been published in Mol. Biol. Evol. [61]. As mentioned earlier, the first part of this work was done for my MSc. But the first year of my PhD was spent extending the work and writing it for publication.

1.6 **Future work**

The simulation study presented in Chapter 3 shows that change in proportion of variable sites, which is not yet incorporated in phylogenetic models, can cause model misspecification and mislead tree reconstruction methods. It is therefore important to design models that account for this feature of the evolutionary process. The simulation tool, LineageSpecificSeqgen, described in Chapter 2 can be used to support development of models for lineage-specific processes which are yet to be modeled, such as changes in the proportion of variable sites. It can be extended to include changes in the substitution process, like those that are identified using the newly developed non-parametric bootstrap test described in Chapter 4. Not only will such models improve phylogenetic tree estimation, but they will also allow us to test our understanding of the underlying molecular evolutionary process. These lineage-specific properties are expected to be present in many molecular datasets; particularly ones for which a long-branch attraction artifact has been suggested as a cause for unexpected phylogenies e.g. Microsporidia. Studying more of these datasets will aid us in understanding the evolutionary process and increase phylogenetic reconstruction accuracy. Simulation studies are widely used in phylogenetics. For these to be useful, it is essential that the processes used to produce the simulated data closely model the underlying biological processes. The difficulties doing this for the case presented in Chapter 5 suggest that there is much room for development in that area. The combination of accounting for lineage-specific evolution and model assessment is expected to be very powerful.

1.7 References

1. Ochman H, Lawrence JG and Groisman EA. 2000. Lateral gene transfer and the nature of bacterial innovation. 405:299-304.
2. Felsenstein J. 2004. Inferring phylogenies. Sinauer Associates, Inc., Sunderland, Massachusetts.
3. Rambaut A and Bromham L. 1998. Estimating divergence dates from molecular sequences. Mol Biol Evol 15:442-448.
4. Drummond AJ, Ho SYW, Phillips MJ and Rambaut A. 2006. Relaxed Phylogenetics and Dating with Confidence. PLoS Biol 4.
5. Thorne JL, Kishino H and Painter IS. 1998. Estimating the rate of evolution of the rate of molecular evolution. Mol Biol Evol 15:1647-1657.
6. Needleman SB and Wunsch CD. 1970. A general method applicable to the search for similarities in the amino acid sequence of two proteins. J Mol Biol 48:443-453.
7. Smith TF and Waterman MS. 1981. Identification of common molecular subsequences. J Mol Biol 147:195-197.
8. Martin W, Roettger M and Lockhart PJ. 2007. A reality check for alignments and trees. Trends Genet 23:478-480.
9. Landan G and Graur D. 2007. Heads or Tails: A Simple Reliability Check for Multiple Sequence Alignments. Mol Biol Evol 24:1380-1383.
10. Huelsenbeck JP, Larget B, Miller RE and Ronquist F. 2002. Potential Applications and Pitfalls of Bayesian Inference of Phylogeny. Syst Biol 51:673-688.
11. Beaumont MA and Rannala B. 2004. The Bayesian revolution in genetics. 5:251-261.

12. Chor B, Hendy MD, Holland BR and Penny D. 2000. Multiple maxima of likelihood in phylogenetic trees: an analytic approach. Mol Biol Evol 17:1529-1541.

13. Holder M and Lewis PO. 2003. Phylogeny estimation: traditional and Bayesian approaches. Nat Rev Genet 4:275-284.

14. Efron B. 1979. Bootstrap Methods: Another Look at the Jackknife. Annals of Statistics 7:1-26.

15. Felsenstein J. 1985. Confidence-Limits on Phylogenies - an Approach Using the Bootstrap. Evolution 39:783-791.

16. Efron B and Tibshirani R. 1993. An introduction to the bootstrap. Chapman&Hall/CRC press.

17. Phillips MJ, Delsuc F and Penny D. 2004. Genome-scale phylogeny and the detection of systematic biases. Mol Biol Evol 21:1455-1458.

18. Simmons MP, Pickett KM and Miya M. 2004. How Meaningful Are Bayesian Support Values? Mol Biol Evol 21:188-199.

19. Lanave C, Preparata G, Saccone C and Serio G. 1984. A new method for calculating evolutionary substitution rates. J Mol Evol 20:86-93.

20. Bos DH and Posada D. 2005. Using models of nucleotide evolution to build phylogenetic trees. Dev Comp Immunol 29:211-227.

21. Jones DT, Taylor WR and Thornton JM. 1992. The rapid generation of mutation data matrices from protein sequences. Comput. Appl. Biosci. 8:275-282.

22. Whelan S and Goldman N. 2001. A General Empirical Model of Protein Evolution Derived from Multiple Protein Families Using a Maximum-Likelihood Approach. Mol Biol Evol 18:691-699.

23. M.O. Dayhoff, R.M. Schwartz and Orcutt BC. 1978. A model of evolutionary change in proteins. Pp. 345-352 *in* Dayhoff MO, ed. Atlas of Protein Sequence Structure. National Biomedical Research Foundation, Washington DC.

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

25. Waddell PJ, Penny D and Moore T. 1997. Hadamard conjugations and modeling sequence evolution with unequal rates across sites. Mol Phylogenet Evol 8:33-50.

26. Fitch WM and Markowitz E. 1970. An Improved Method for Determining Codon Variability in a Gene and Its Application to Rate of Fixation of Mutations in Evolution. Biochem Genet 4:579-593.

27. Tuffley C and Steel M. 1998. Modeling the covarion hypothesis of nucleotide substitution. Math Biosci 147:63-91.

28. Huelsenbeck JP. 2002. Testing a covariotide model of DNA substitution. Mol Biol Evol 19:698-707.

29. Galtier N. 2001. Maximum-likelihood phylogenetic analysis under a covarion-like model. Mol Biol Evol 18:866-873.

30. Wang HC, Spencer M, Susko E and Roger AJ. 2007. Testing for covarion-like evolution in protein sequences. Mol Biol Evol 24:294-305.

31. Bromham LD. 2001. Molecular Evolution: Rates. Encyclopedia of Life Science. Macmillian Publishers Ltd.

32. Philippe H, Casane D, Gribaldo S, Lopez P and Meunier J. 2003. Heterotachy and functional shift in protein evolution. IUBMB Life 55:257-265.

33. Fraser HB, Wall DP and Hirsh AE. 2003. A simple dependence between protein evolution rate and the number of protein-protein interactions. BMC Evol Biol 3.

34. Saeed R and Deane CM. 2006. Protein protein interactions, evolutionary rate, abundance and age. BMC Bioinformatics 7.

35. Lopez P, Casane D and Philippe H. 2002. Heterotachy, an important process of protein evolution. Mol Biol Evol 19:1-7.

36. Dickerson RE. 1971. The structures of cytochrome c and the rates of molecular evolution. Molecular Evolution 1:26-45.

37. Ane C, Burleigh JG, McMahon MM and Sanderson MJ. 2005. Covarion structure in plastid genome evolution: A new statistical test. Mol Biol Evol 22:914-924.

38. Lockhart P, Novis P, Milligan BG, Riden J, Rambaut A and Larkum T. 2006. Heterotachy and tree building: A case study with plastids and eubacteria. Mol Biol Evol 23:40-45.

39. Steel M and Penny D. 2000. Parsimony, likelihood, and the role of models in molecular phylogenetics. Mol Biol Evol 17:839-850.

40. Sullivan J and Joyce P. 2005. Model selection in phylogenetics. Annu Rev Ecol Evol Syst 36:445-466.

41. Matsen FA and Steel M. 2007. Phylogenetic Mixtures on a Single Tree Can Mimic a Tree of Another Topology. Syst Biol 56:767-775.

42. Kelchner SA and Thomas MA. 2007. Model use in phylogenetics: nine key questions. Trends Ecol Evol 22:87-94.

43. Akaike H. 1974. A new look at the statistical model identification. Automatic Control, IEEE Transactions on 19:716-723.

44. Schwarz G. 1978. Estimating the Dimension of a Model. The Annals of Statistics 6:461-464.

45. Bollback JP. 2002. Bayesian model adequacy and choice in phylogenetics. Mol Biol Evol 19:1171-1180.

46. Goldman N. 1993. Statistical Tests of Models of DNA Substitution. J Mol Evol 36:182-198.

47. Posada D and Buckley TR. 2004. Model selection and model averaging in phylogenetics: advantages of akaike information criterion and bayesian approaches over likelihood ratio tests. Syst Biol 53:793-808.

48. Brown JM and ElDabaje R. 2009. PuMA: Bayesian analysis of partitioned (and unpartitioned) model adequacy. Bioinformatics 25:537-538.

49. Felsenstein J. 1978. Cases in Which Parsimony or Compatibility Methods Will Be Positively Misleading. Syst Zool 27:401-410.

50. Hendy MD and Penny D. 1989. A Framework for the Quantitative Study of Evolutionary Trees. Syst Zool 38:297-309.

51. Bergsten J. 2005. A review of long-branch attraction. Cladistics 21:163-193.

52. Philippe H and Laurent J. 1998. How good are deep phylogenetic trees? Curr Opin Genet Dev 8:616-623.

53. Holland BR, Penny D and Hendy MD. 2003. Outgroup misplacement and phylogenetic inaccuracy under a molecular clock - A simulation study. Syst Biol 52:229-238.

54. Inagaki Y, Susko E, Fast NM and Roger AJ. 2004. Covarion shifts cause a long-branch attraction artifact that unites microsporidia and archaebacteria in EF-1 alpha phylogenies. Mol Biol Evol 21:1340-1349.

55. Soltis DE, Albert VA, Savolainen V, Hilu K, Qiu YL, Chase MW, Farris JS, Stefanovic S, Rice DW, Palmer JD and Soltis PS. 2004. Genome-scale data, angiosperm relationships, and 'ending incongruence': a cautionary tale in phylogenetics. Trends Plant Sci 9:477-483.

56. Liu Y, Leigh JW, Brinkmann H, Cushion MT, Rodriguez-Ezpeleta N, Philippe H and Lang BF. 2009. Phylogenomic Analyses Support the Monophyly of Taphrinomycotina, including Schizosaccharomyces Fission Yeasts. Mol Biol Evol 26:27-34.

57. Brinkmann H, van der Giezen M, Zhou Y, Poncelin de Raucourt G and Philippe H. 2005. An empirical assessment of long-branch attraction artefacts in deep eukaryotic phylogenomics. Syst Biol 54:743-757.

58. Yang Z and Nielsen R. 2002. Codon-Substitution Models for Detecting Molecular Adaptation at Individual Sites Along Specific Lineages. Mol Biol Evol 19:908-917.

59. Guindon Sp, Rodrigo AG, Dyer KA and Huelsenbeck JP. 2004. Modeling the site-specific variation of selection patterns along lineages. Proc Natl Acad Sci U S A 101:12957-12962.

60. Hillis DM, Huelsenbeck JP and Cunningham CW. 1994. Application and Accuracy of Molecular Phylogenies. Science 264:671-677.

61. Shavit L, Penny D, Hendy MD and Holland BR. 2007. The Problem of Rooting Rapid Radiations. Mol Biol Evol 24:2400-2411.

62. Gruenheit N, Lockhart PJ, Steel M and Martin W. 2008. Difficulties in testing for covarion-like properties of sequences under the confounding influence of changing proportions of variable sites. Mol Biol Evol 25:1512-1520.

63. Kolaczkowski B and Thornton JW. 2004. Performance of maximum parsimony and likelihood phylogenetics when evolution is heterogeneous. Nature 431:980-984.

64. Ruano-Rubio V and Fares MA. 2007. Artifactual phylogenies caused by correlated distribution of substitution rates among sites and lineages: the good, the bad, and the ugly. Syst Biol 56:68-82.

65. Pagel M and Meade A. 2004. A phylogenetic mixture model for detecting pattern-heterogeneity in gene sequence or character-state data. Syst Biol 53:571-581.

66. Soria-Carrasco V, Talavera G, Igea J and Castresana J. 2007. The K tree score: quantification of differences in the relative branch length and topology of phylogenetic trees. Bioinformatics 23:2954-2956.

67. Beaumont MA, Zhang W and Balding DJ. 2002. Approximate Bayesian computation in population genetics. Genetics 162:2025-2035.

68. Shavit Grievink L, Penny D, Hendy MD and Holland BR. 2008. LineageSpecificSeqgen: generating sequence data with lineage-specific variation in the proportion of variable sites. BMC Evol Biol 8:317.

69. Kolaczkowski B and Thornton JW. 2008. A mixed branch length model of heterotachy improves phylogenetic accuracy. Mol Biol Evol.

70. Steel M. 2005. Should phylogenetic models be trying to 'fit an elephant? Trends Genet 21:307-309.

71. Philippe H, Zhou Y, Brinkmann H, Rodrigue N and Delsuc F. 2005. Heterotachy and long-branch attraction in phylogenetics. BMC Evol Biol 5.

72. Spencer M, Susko E and Roger AJ. 2005. Likelihood, parsimony, and heterogeneous evolution. Mol Biol Evol 22:1161-1164.

73. Gadagkar SR and Kumar S. 2005. Maximum likelihood outperforms maximum parsimony even when evolutionary rates are heterotachous. Mol Biol Evol 22:2139-2141.

Chapter 2

LineageSpecificSeqgen: generating sequence data with lineage-specific variation in the proportion of variable sites

As published in BMC Evolutionary Biology.

Ref: Shavit Grievink, L., D. Penny, M. D. Hendy, and B. R. Holland. 2008. LineageSpecificSeqgen: generating sequence data with lineage-specific variation in the proportion of variable sites. BMC Evol Biol 8:317.

Shavit Grievink, L., D. Penny, M. D. Hendy, and B. R. Holland. 2009. Correction: LineageSpecificSeqgen: generating sequence data with lineage-specific variation in the proportion of variable sites. Submitted to BMC Evolutionary Biology.

2.1 Abstract

2.1.1 Background

Commonly used phylogenetic models assume a homogeneous evolutionary process throughout the tree. It is known that these homogeneous models are often too simplistic, and that with time some properties of the evolutionary process can change (due to selection or drift). In particular, as constraints on sequences evolve, the proportion of variable sites can vary between lineages. This affects the ability of phylogenetic methods to correctly estimate phylogenetic trees, especially for long timescales. To date there is no phylogenetic model that allows for change in the proportion of variable sites, and the degree to which this affects phylogenetic reconstruction is unknown.

2.1.2 Results

We present LineageSpecificSeqgen, an extension to the seq-gen program that allows generation of sequences with both changes in the proportion of variable sites and changes in the rate at which sites switch between being variable and invariable. In contrast to seq-gen and its derivatives to date, we interpret branch lengths as the mean number of substitutions per variable site, as opposed to the mean number of substitutions per site (which is averaged over all sites, including invariable sites). This allows specification of the substitution rates of variable sites, independently of the proportion of invariable sites.

2.1.3 Conclusions

LineageSpecificSeqgen allows simulation of DNA and amino acid sequence alignments under a lineage-specific evolutionary process. The program can be used to test current models of evolution on sequences that have undergone lineage-specific evolution. It facilitates the development of both new methods to identify such processes in real data,

and means to account for such processes. The program is available at: http://awcmee.massey.ac.nz/downloads.htm.

2.2 **Background**

Simulated sequence data are widely used for hypothesis testing [1], for evaluation of phylogenetic methods under different parameter settings [2, 3, 4], for testing the effect of model misspecification on tree reconstruction [5, 6, 7], for development of new models and methods [8, 9], and for approximate Bayesian inference [10]. For these applications, it is important that the processes used to produce the simulated data closely model the underlying biological processes. Commonly used phylogenetic sequence generators employ homogeneous, time reversible, stationary models of molecular sequence evolution. These phylogenetic models assume that the overall rate of substitution is the only parameter that may change along the tree and do not allow changes in other parameters, such as the rate matrix, the distribution of rates across sites and the proportion of variable sites.

It is known, however, that as sequences diverge they can acquire independent properties. In particular, the proportion of variable sites can evolve in a lineage-specific manner due to changes in evolutionary constraints [11, 12]. The proportion of variable sites in a lineage will affect its estimated substitution rate [13]. Failure to account for changes in the proportion of variable sites can result in erroneous rate estimates that may affect tree estimation [5, 14]. Indeed, change in the proportion of variable sites is thought to be one of the main causes of long-branch attraction [12, 15].

In addition to the possible shift in the proportion of variable sites, it is known that sites can switch between variable and invariable states due to drift. Note that invariable sites (which we are concerned with in this paper) are sites for which the probability of character substitution is zero; as opposed to invariant sites for which the probability of character substitution is greater than zero but for a certain group (sample) of taxa no substitution is found. The strict covarion model [16] allows sites to switch between variable and invariable states; however, at equilibrium the proportion of variable sites is constant over the different lineages. Several extensions of the covarion model [17, 18,

19] are implemented in the sequence generator seq-gen-aminocov [19]. However, to date, there is no model that allows for change in the proportion of variable sites.

Using partitions (a partition is a group of consecutive sites that are simulated on the same underlying tree), lineage-specific proportions of variable sites have been simulated with sequence generators such as seq-gen [20], seq-gen-cov [21], and seq-gen-aminocov [19]. These simulations have proven to be very useful in facilitating our understanding of the process of lineage-specific evolution. However, the use of these programs for the purpose of simulating changes in the proportion of variable sites is limited to trees with very few 'events', where an event is defined as a position on the tree where a change in the process of evolution occurs, e.g. a change in the proportion of variable sites. This is because different proportions of variable sites are generated using pre-defined partitions, where each partition is simulated on a tree with different branch lengths (zero branch lengths are used for invariable sites). For two events, in which the proportion of variable sites changes, there are 8 partitions [5]. In general there are $2^{(1+\text{number of events})}$ partitions (M. Steel, personal communication), so creating the input for such simulations becomes a difficult task.

Furthermore, in seq-gen, invariable sites can be incorporated into sequences by either simulating on different partitions (where a partition for invariable sites is simulated on a zero length tree), or specifying a proportion of invariable sites (Pinv) using the –i option. Intuitively, one might expect the processes of evolution simulated by these two methods to be equivalent, but this is not the case. In seq-gen and its modifications published to date, branch lengths are defined as the mean expected number of substitutions per site. When sequences are simulated with a specified proportion of invariable sites, the branch lengths specified by the user are rescaled (increased) by the program to compensate for the proportion of invariable sites. Hence, increasing the proportion of invariable sites (for which the substitution rate is zero) forces a greater substitution rate on the variable sites. For example, with 80% invariable sites and an expected mean number of substitutions of 0.02, the mean number of substitutions of the variable sites will be rescaled to 0.1. Although this branch rescaling is consistent with

the definition of branch lengths as the mean expected number of substitutions per site, we found that many researchers are not aware of it.

Moreover, using partitions does not allow changes in the on/off switch rate of the covarion model. We have developed a program that allows the user to simulate sequence data containing changes in the proportion of variable sites, and changes in the covarion switch rate, without the need to specify partitions or rescale branch lengths.

2.3 **Implementation**

LineageSpecificSeqgen is a command-line controlled program written in C. The program uses, as much as possible, the code from seq-gen and its derivatives [19, 20, 21]. Given a rooted tree, specified events (in which changes in the process occur), and a set of parameters, the program generates a user-specified number of datasets (nucleotide or amino-acid). An example workflow of the program is illustrated in Figure 2.1. The input is two text files - a tree file and a parameter file. The tree file contains one or more trees in a format which is based on the Newick format. Events on the tree are marked using a $ sign and are given names. Lengths are specified for all branches of the tree; for branches with events, the length before and the length after the event must both be specified. The parameter file contains the changes in the proportion of variable sites, and/or the switch rate of the covarion process, at each event. Any number of events can be specified. A change in the proportion of variable sites is specified using two parameters; the proportion of sites that were invariable and became variable at the event, and the proportion of sites that were variable and became invariable at the event.

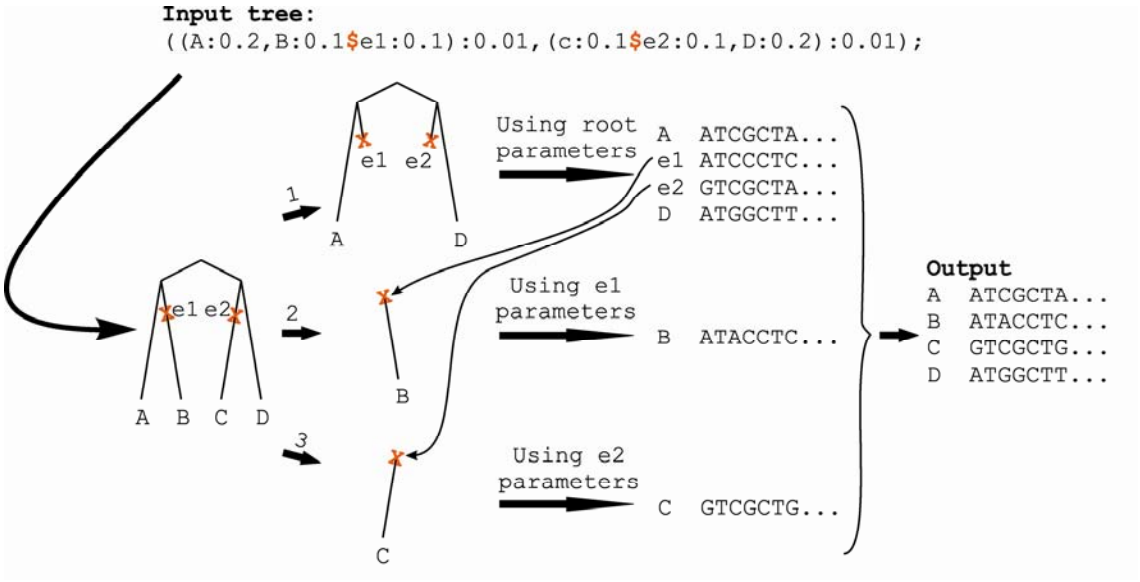


Figure 2.1 - An example workflow of LineageSpecificSeqgen. For each tree, the program creates a random root sequence according to the parameters specified. The program then evolves the sequences, according to the parameters given for the root, along the subtree beneath the root (excluding parts of the tree that are beneath events). The resulting sequence at each event is then used as an ancestral sequence for the subtree beneath that event, and the sequences are evolved along that subtree according to the parameters specified for that event. The output is an alignment of the resulting sequences at the tips.

For each input tree, the program will generate *n* subtrees (*n* = 1+number of events), with each event on the tree defining a cutting point (see Figure 2.1). For each input tree and each dataset, sequences are first simulated on the subtree under the root and then on the subtree under each event in an iterative manner. An array holding the state (variable/invariable) of each position is updated at each event according to the change in the proportion of variable sites specified by the user. Events can be specified as correlated, although by default they are non-correlated. For correlated events the positions of sites that switch state are identical, for non-correlated events these positions are independent. An array holding the hidden states of the covarion model is also passed down the tree. For each site, along each branch, exponential times for switches are generated; the hidden states array is updated at the internal nodes of each subtree according to the specified covarion model and the switch rate for each event. The sequence at each event is used as the ancestral sequence for the subtree beneath it. The output is an alignment of the resulting tip sequences.

For the reasons described in the Background, we added a default option where branch lengths are defined as the mean expected numbers of substitutions per variable site. This definition allows the substitution rate across variable sites to be independent of the proportion of invariable sites. When branch lengths are defined as the mean expected numbers of substitutions per variable site, the processes of evolution simulated by both specified partitions and specified Pinv are equivalent.

2.4 Results And Discussion

2.4.1 Example 1 – generating data containing a change in the proportion of variable sites

To demonstrate the use of the program for generating datasets containing a change in the proportion of variable sites, the Jukes-Cantor (JC) model was used to generate sequences of length 10,000bp on the 16-taxon rooted balanced tree shown in Figure 2.2.

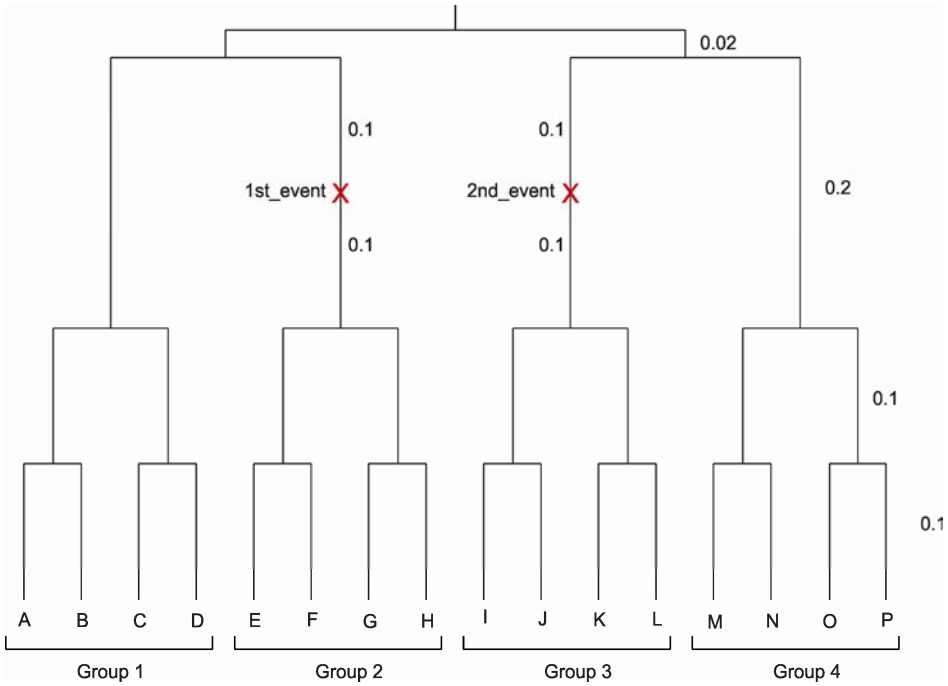


Figure 2.2 - 16-taxon rooted balanced tree used for simulation. Sequences were generated on a 16-taxon rooted balanced tree. The tree is comprised of four groups of four taxa each. There are two correlated events on the tree in which the proportion of variable sites changes. The events are located on the two non-sister lineages 2 and 3.

This tree is input as:

((((A:0.1,B:0.1):0.1,(C:0.1,D:0.1):0.1):0.2,((E:0.1,F:0.1):0.1,(G:0.1,H:0.1):0.1):0.1$1st_event:0.1):0.02,(((I:0.1,J:0.1):0.1,(K:0.1,L:0.1):0.1):0.1$2nd_event:0.1,((M:0.1,N:0.1):0.1,(O:0.1,P:0.1):0.1):0.2):0.02);

In this example, the proportion of invariable sites (Pinv) at the root was set to 0.8. At each of the two events 0.2 of the invariable sites were "switched on" (became variable). The two events were set to be correlated so that the positions of sites that are turned on in the two events are identical. The expected proportion of variable sites in groups 1 and 4 is thus 0.2, and the expected proportion of variable sites in groups 2 and 3 is 0.36. Consequentially, 0.64 of the sites are invariable across all four groups, and 0.16 of the sites are variable in groups 2 and 3 and invariable in groups 1 and 4. For comparison, two control datasets were generated on a 16-taxon rooted balanced tree without the two events; the same branch lengths were used as before, and Pinv was set to either 0 or 0.8. For each group, and each pair of groups, the number of sites that varied in each of the three simulated sequence alignments is shown in Table 2.1.

Table 2.1 - Number of sites that vary in each of the four groups, and each pair of groups, for the three datasets.

Group/s	No events Pinv=0 (all sites are variable)	No events Pinv=0.8	Pinv=0.8 Two correlated events Pvar+=0.2
1	4509	905	911
2	4363	922	1574
3	4347	925	1635
4	4410	913	919
1 and 2	1947	404	375
1 and 3	1930	410	416
1 and 4	2003	411	417
2 and 3	1915	426	709
2 and 4	1901	431	409
3 and 4	1898	421	408

2.4.2 Example 2 – testing tree reconstruction accuracy for data containing a change in the proportion of variable sites

To demonstrate the use of the program for testing tree reconstruction accuracy for datasets containing a change in the proportion of variable sites, the JC model was used

to generate 100 datasets of length 10,000bp on the 4-taxon rooted balanced tree shown in Figure 2.3.

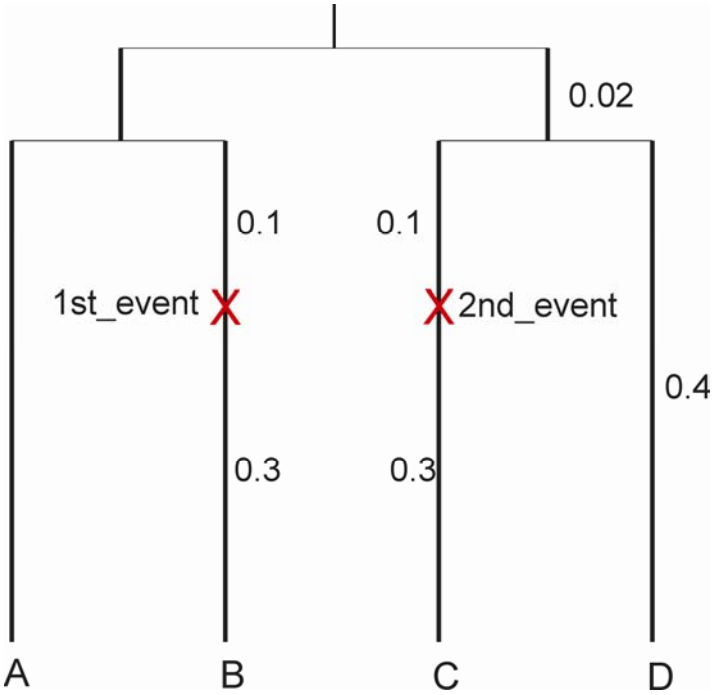


Figure 2.3 - 4-taxon rooted balanced tree used for simulation. Sequences were generated on a 4-taxon rooted balanced tree. There are two correlated events on the tree in which the proportion of variable sites changes. The events are located on the two non-sister lineages 2 and 3.

This tree is input as:
((A:0.4,B:0.3$1st_event:0.1):0.02,(C:0.3$2nd_event:0.1,D:0.4):0.02);

As in the former example, Pinv at the root was set to 0.8 and the two events were set to be correlated. At each of the two events $Pvar^{+}$ = (0,5,10,15,20,25,30) percent of the invariable sites were "switched on". The program MrBayes [22] was used to reconstruct the trees, assuming a JC model with invariable sites and a covarion process (JC+I+covarion). The number of times with which each of the three possible 4-taxon trees was reconstructed with the highest proportional frequency in the Bayesian analysis were compared to determine tree reconstruction accuracy. As shown in Figure 2.4, the higher the increase in the proportion of variable sites in lineages B and C, the lower the tree reconstruction accuracy. These results suggest that, at least for some parts of the parameter space, a covarion model which assumes a constant proportion of variable

sites is not adequate for tree reconstruction from data containing changes in the proportion of variable sites.

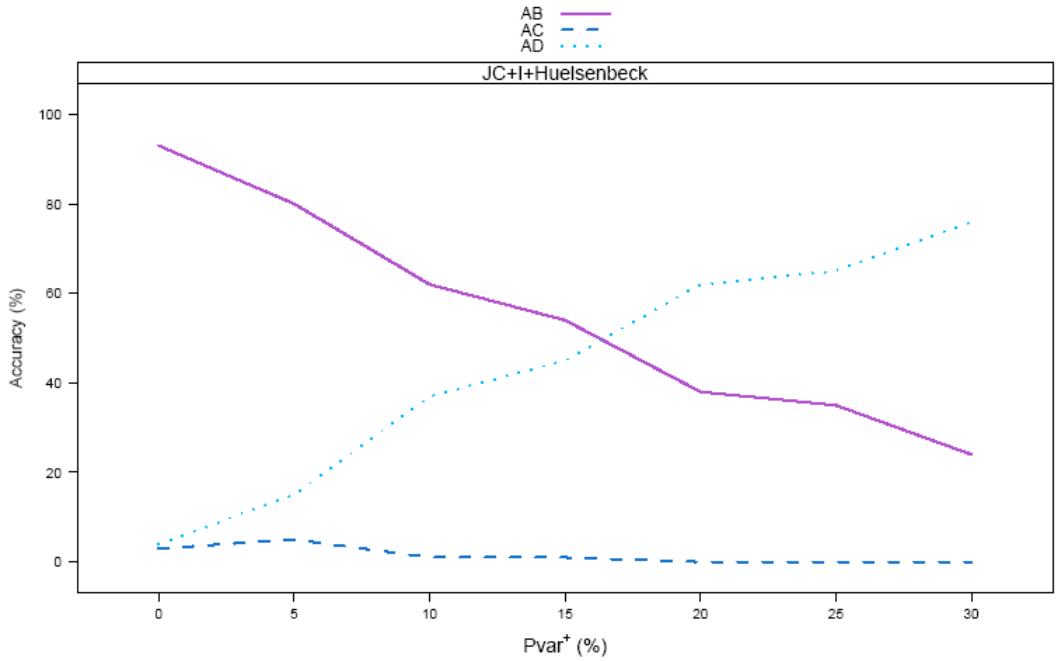


Figure 2.4 - Tree reconstruction accuracy for the simulated 4-taxon datasets, using JC+I+covarion model. The number of times with which each of the three possible 4-taxon trees was reconstructed, with the highest proportional frequency in the Bayesian analysis, assuming JC+I+covarion model. The higher the increase in the proportion of variable sites in lineages B and C, the lower the tree reconstruction accuracy.

2.5 Conclusions

LineageSpecificSeqgen is a sequence generator that allows simulation of changes in the proportion of variable sites, a biochemically realistic process of evolution. It is useful for testing current models of evolution on sequences that have undergone lineage-specific evolution, developing methods to identify such processes in real data, and developing means to account for such processes.

2.6 Availability and requirements

- Project name: LineageSpecificSeqgen
- LineageSpecificSeqgen, including the source code and documentation, can be downloaded from http://awcmee.massey.ac.nz/downloads.htm.
- Operating System: The program can be compiled and run on Unix, Linux, and Mac OS.
- Programming Language: ANSI C.
- Other requirements: None.
- License: GNU GPL.
- Any restrictions to use by non-academics: None.
- LineageSpecificSeqgen is provided with no guarantee or warranty of any kind, although the authors are happy to provide assistance if needed.

2.7 **Authors' contributions**

LSG has developed and implemented the code for the program, and had written the first draft of this manuscript. DP, MDH and BRH supervised the project. All the authors contributed to the writing of this manuscript.

2.8 **Acknowledgements**

We thank Pete Lockhart, Bill Martin and Mike Steel for their invaluable contribution in discussions. We also thank Matthew Spencer for his suggestions for debugging.
This work was financially supported by the New Zealand Marsden fund (05-MAU-033 to B.R.H).

2.9 References

1. Goldman N. 1993a. Statistical Tests of Models of DNA Substitution. J Mol Evol 36:182-198.
2. Hillis DM, Huelsenbeck JP and Cunningham CW. 1994. Application and Accuracy of Molecular Phylogenies. Science 264:671-677.
3. Holland BR, Penny D and Hendy MD. 2003. Outgroup misplacement and phylogenetic inaccuracy under a molecular clock - A simulation study. Syst Biol 52:229-238.
4. Shavit L, Penny D, Hendy MD and Holland BR. 2007. The Problem of Rooting Rapid Radiations. Mol Biol Evol 24:2400-2411.
5. Gruenheit N, Lockhart PJ, Steel M and Martin W. 2008. Difficulties in testing for covarion-like properties of sequences under the confounding influence of changing proportions of variable sites. Mol Biol Evol 25:1512-1520.
6. Kolaczkowski B and Thornton JW. 2008. A mixed branch length model of heterotachy improves phylogenetic accuracy. Mol Biol Evol.
7. Ruano-Rubio V and Fares MA. 2007. Artifactual phylogenies caused by correlated distribution of substitution rates among sites and lineages: the good, the bad, and the ugly. Syst Biol 56:68-82.
8. Pagel M and Meade A. 2004. A phylogenetic mixture model for detecting pattern-heterogeneity in gene sequence or character-state data. Syst Biol 53:571-581.
9. Soria-Carrasco V, Talavera G, Igea J and Castresana J. 2007. The K tree score: quantification of differences in the relative branch length and topology of phylogenetic trees. Bioinformatics 23:2954-2956.
10. Beaumont MA, Zhang W and Balding DJ. 2002. Approximate Bayesian computation in population genetics. Genetics 162:2025-2035.
11. Lopez P, Casane D and Philippe H. 2002. Heterotachy, an important process of protein evolution. Mol Biol Evol 19:1-7.

12. Lockhart P, Novis P, Milligan BG, Riden J, Rambaut A and Larkum T. 2006. Heterotachy and tree building: A case study with plastids and eubacteria. Mol Biol Evol 23:40-45.

13. Dickerson RE. 1971. The structures of cytochrome c and the rates of molecular evolution. Molecular Evolution 1:26-45.

14. Lockhart PJ and Steel MA. 2005. A Tale of Two Processes. Syst Biol 54:948-951.

15. Inagaki Y, Susko E, Fast NM and Roger AJ. 2004. Covarion shifts cause a long-branch attraction artifact that unites microsporidia and archaebacteria in EF-1 alpha phylogenies. Mol Biol Evol 21:1340-1349.

16. Tuffley C and Steel M. 1998. Modeling the covarion hypothesis of nucleotide substitution. Math Biosci 147:63-91.

17. Galtier N. 2001. Maximum-likelihood phylogenetic analysis under a covarion-like model. Mol Biol Evol 18:866-873.

18. Huelsenbeck JP. 2002. Testing a covariotide model of DNA substitution. Mol Biol Evol 19:698-707.

19. Wang HC, Spencer M, Susko E and Roger AJ. 2007. Testing for covarion-like evolution in protein sequences. Mol Biol Evol 24:294-305.

20. Rambaut A and Grassly NC. 1997. Seq-Gen: An application for the Monte Carlo simulation of DNA sequence evolution along phylogenetic frees. Computer Applications in the Biosciences 13:235-238.

21. Ane C, Burleigh JG, McMahon MM and Sanderson MJ. 2005. Covarion structure in plastid genome evolution: A new statistical test. Mol Biol Evol 22:914-924.

22. Ronquist F and Huelsenbeck JP. 2003. MrBayes 3: Bayesian phylogenetic inference under mixed models. Bioinformatics 19:1572-1574.

2.10 **Correction**

Since publication of our article [1], we discovered an error in the second example. For this example, we state in the paper, we used the program MrBayes [2] with the JC+I+Cov model. However, we now found that, albeit appearances, this model is not implemented in MrBayes [2]. In fact, no combination of I+Cov (e.g. HKY+I+Cov, GTR+G+I+Cov) is currently implemented in MrBayes [2]. Instead, the program ignores the I parameter, so tree reconstruction in this example was therefore effectively done using the JC+Cov model. This does not affect the conclusion of our paper that phylogenetic estimation can be misleading for sequence data simulated with lineage-specific properties.

2.10.1 References

1. Shavit Grievink L, Penny D, Hendy MD and Holland BR. 2008. LineageSpecificSeqgen: generating sequence data with lineage-specific variation in the proportion of variable sites. BMC Evol Biol 8:317.
2. Ronquist F and Huelsenbeck JP. 2003. MrBayes 3: Bayesian phylogenetic inference under mixed models. Bioinformatics 19:1572-1574.

Chapter 3

Phylogenetic tree reconstruction accuracy and model fit when proportions of variable sites change across the tree

Manuscript; accepted for publication, pending revisions, in Systematic Biology.

(Note: most suggested revisions have been incorporated in this chapter)

Ref: Shavit Grievink, L., D. Penny, M. D. Hendy, and B. R. Holland. 2009. Phylogenetic tree reconstruction accuracy and model fit when proportions of variable sites change across the tree. Syst Biol.

3.1 **Abstract**

Commonly used phylogenetic models assume a homogeneous process through time in all parts of the tree. However, it is known that these models can be too simplistic as they do not account for non-homogeneous lineage-specific properties. In particular, it is now widely recognized that as constraints on sequences evolve the proportion and positions of variable sites can vary between lineages causing heterotachy. The extent to which this model misspecification affects tree reconstruction is still unknown. Here, we evaluate the effect of changes in the proportions and positions of variable sites on model fit and tree estimation. We consider five current models of nucleotide sequence evolution in a Bayesian MCMC framework, as well as maximum parsimony. We show that for a tree with 4 lineages where 2 non-sister taxa undergo a change in the proportion of variable sites tree reconstruction under the best-fitting model, which is chosen using a relative test, often results in the wrong tree. In this case we found that an absolute test of model-fit is a better predictor of tree estimation accuracy. We also found further evidence that maximum-parsimony is not immune to heterotachy. In addition, we show that increased sampling of taxa that have undergone a change in proportion and positions of variable sites is critical for accurate tree reconstruction.

Key Words: [Phylogenetics, Heterotachy, Covarion model, Model fit, Taxon sampling, Simulation]

3.2 **Introduction**

Commonly used phylogenetic models assume a homogeneous, time reversible, stationary process, at each site, throughout the tree. However, it is known that these assumptions are a simplification of the true evolutionary process. In particular, a site can display lineage-specific rates of substitution, an observation that has been termed heterotachy [1]. This type of variation appears to be a prevalent feature of molecular sequence data [2, 3, 4, 5]; however some evolutionary processes that can cause heterotachy are not accounted for in phylogenetic models. Such model misspecification can mislead model-based tree reconstruction [6, 7].

Heterotachy arises from different evolutionary processes including changes in (1) the overall rates of substitutions, (2) the positions of variable sites, and/or (3) the proportions of variable sites. These processes are likely to be correlated and reflect variations, over time, in the underlying evolutionary constraints that are acting on the sequences. Importantly, the latter two processes, which can be explained biochemically by changes in the evolutionary constraints acting on the secondary and tertiary structures, can explain the observed changes in overall rates as well as variations in rates across sites.

Here we focus on changes in the proportions and positions of variable sites, and their effect on model fit and tree reconstruction. Although such changes are known to occur over time independently in different lineages [2, 3, 4, 5, 8] and have been shown to mislead tree reconstruction [6, 9, 10], the extent of their effects on phylogenetic reconstruction is still uncertain. Using simulated data, we measured and compared tree reconstruction accuracies of five current models of nucleotide sequence evolution in a Bayesian MCMC framework, as well as the accuracy of maximum parsimony (MP), when applied to data containing increasing levels of change in the proportions of variable sites (Pvar) with and without additional changing positions of variable sites.

We explore the effect of taxon sampling on the estimation of the inner-most branch. The number of possible trees increases super-exponentially with the number of taxa. Therefore, phylogenetic analysis using a large number of taxa is computationally difficult. However, the inclusion of appropriate additional taxa has previously been found to increase the reconstruction accuracy of underlying relationships particularly when the additional taxa break up long branches [11, 12].

We also examine the relative and absolute adequacy of these models for such data. It is important to note that the best-fit model is not necessarily adequate for tree reconstruction [13]. Model selection methods chose a model, from a given set of models, that maximizes the likelihood of the data given the tree (considering, and in some cases penalizing for the number of parameters). Model adequacy assessment methods (such as [14, 15]) evaluate how well a certain model performs in predicting future observations. This is usually done by simulating predictive observations under the model in question, and comparing these to the original data using some test statistic. Unlike model selection methods, these evaluate the absolute adequacy of the model and can reject the best-fit model if some component of the evolutionary process is not accounted for in the set of models tested [16].

3.3 Material and Methods

3.3.1 Simulations

We generated data using our newly developed simulator LineageSpecificSeqgen [17] (Chapter 2 in this thesis); an extension to the seq-gen program [18] that allows generation of sequences with both changes in the proportions of variable sites (Pvar) and changes in the variable/invariable switch rate of the covarion model [19]. One hundred DNA datasets of 10,000 nucleotides each were generated along the 4-, 6-, 8-, and 16-taxon trees depicted in Figure 3.1. We used the default option of LineageSpecificSeqgen where branch lengths are defined as the expected number of substitutions per variable site; as opposed to the expected number of substitutions per site (which is averaged over all sites, including invariable sites). The advantage of this setting is that it is more intuitive; the input branch lengths are used directly and the rate of variable sites is not increased (rescaled) to compensate for the invariable sites when the data is generated. This results in simulation of more moderate rates than in the alternative setting of branch lengths being the expected number of substitutions per site (see [17], Chapter 2 in this thesis, for further detail). The setting used does not affect tree estimation as the expected number of substitution per site will be estimated from the data.

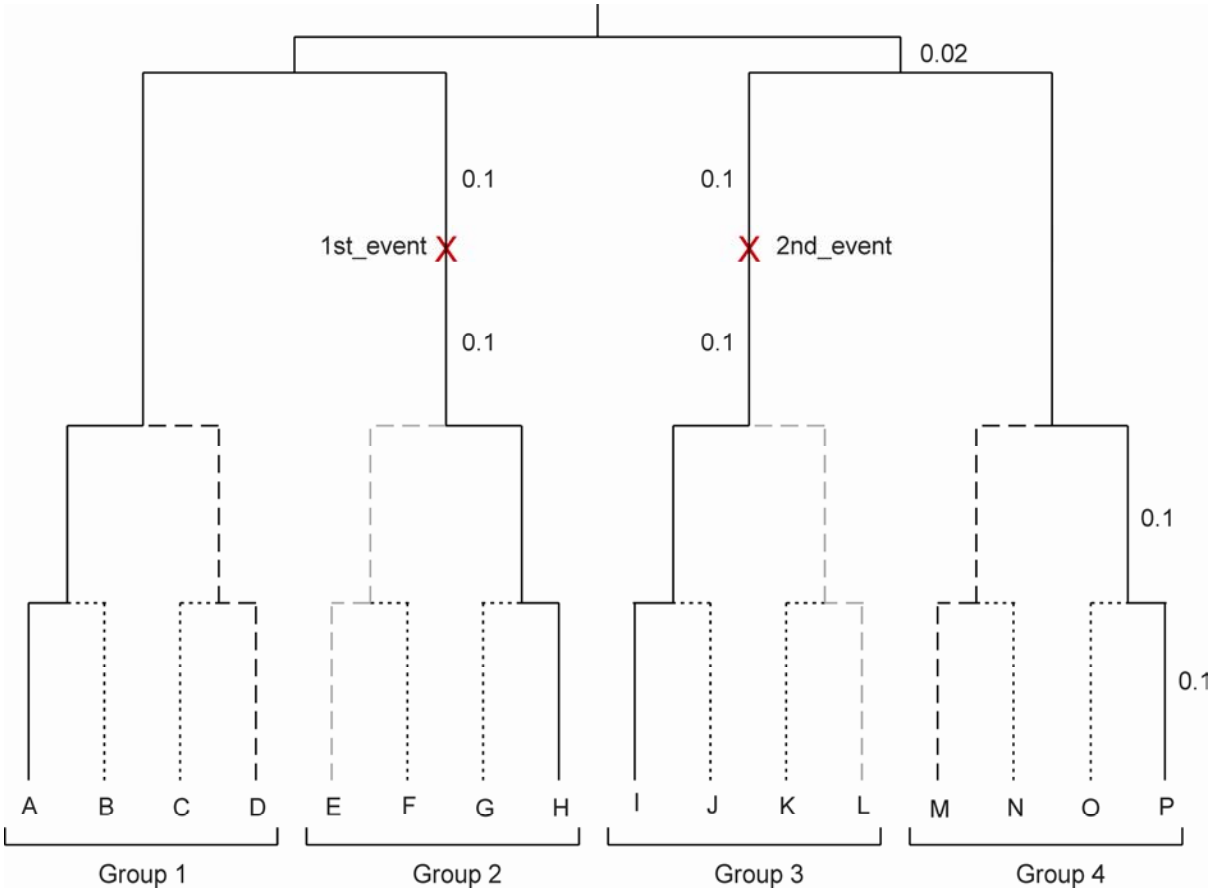


Figure 3.1 - Simulations were done on a 4-taxon tree: T_4=((A,H),(I,P)) [solid lines], two 6-taxon trees: T_{6a}=((A,(E,H)),((I,L),P)) [solid and light dashed lines] and T_{6b}=(((A,D),H),(I,(M,P))) [solid and dark dashed lines], an 8-taxon tree T_8=((A,D),(E,H)),((I,L),(M,P))) [solid and both dashed lines], and a 16-taxon tree

T_{16}=(((A,B),(C,D)),((E,F),(G,H))),(((I,J),(K,L)),((M,N),(O,P)))) [all lines].

The Jukes-Cantor model [20] of nucleotide substitution was used both with and without the covarion model of Tuffley and Steel [19]; the proportion of sites that are variable under the covarion model was set to 0.6 and the rate of change from variable to invariable and vice versa was set to 0.1). As illustrated in Figure 3.2, a site can be invariable at a certain section of the tree if a) it is part of the proportion of sites that are invariable (Pinv) or b) it is part of the proportion of sites that are variable (Pvar) but is invariable ('off') under the covarion model. At the root 80% of the sites were set as invariable (i.e. Pinv=0.8, Pvar=0.2). Changes in the proportion of variable sites (Pvar), 'events', were introduced in two positions on the trees marked as '1st_event' and '2nd_event' (Figure 3.1); $Pvar^+$ = (0, 5, 10, 15, 20, 25, 30, 35, 40, 45, 50) percent of the

invariable sites were reset to be variable in two events. Unless otherwise stated, these two events were set to be correlated, so that the positions of sites that switch state are identical.

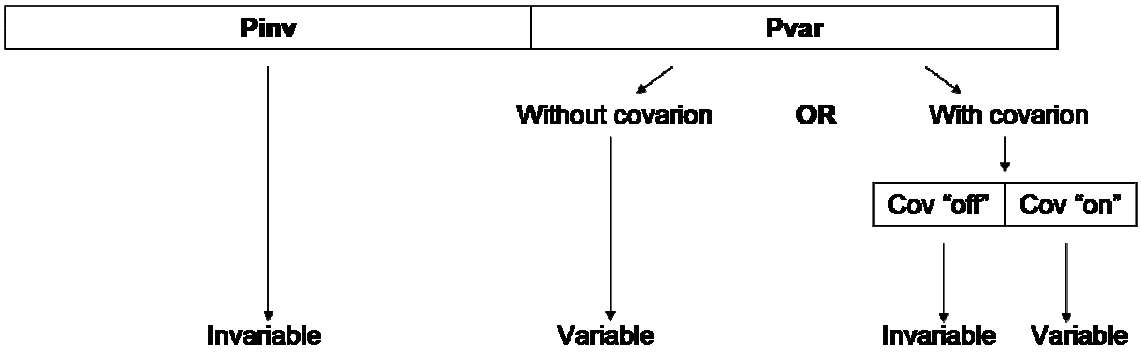


Figure 3.2 - A description of the variable and invariable sites in the simulated data. When sequences are simulated without the covarion model, the number of variable sites is equal to the proportion of variable sites (Pvar) multiplied by the number of sites and thus the number of invariable sites is equal to the proportion of invariable sites (Pinv) multiplied by the number of sites. However, when sequences are simulated with the covarion model, the number of variable sites is equal to the proportion of variable sites (Pvar) multiplied by the proportion of sites that are "on" (i.e. variable) under the covarion model (Cov "on") and the number of sites; the number of invariable sites is then equal to the proportion of invariable sites (Pinv) multiplied by the number of sites plus the proportion of variable sites (Pvar) multiplied by the proportion of sites that are "off" (i.e. invariable) under the covarion model (Cov "off") and the number of sites. A site can therefore be invariable at a certain time if a) it is part of Pinv or b) it is part of Cov "off".

3.3.2 Phylogenetic Analyses

For each simulated dataset, we conducted a Bayesian analysis using MrBayes version 3.1 [21] under five different models: JC, JC with invariable sites (JC+I), JC with a gamma distribution of rates across sites (JC+G), JC with invariable sites and a gamma distribution (JC+I+G), and JC with the covarion model (JC+Cov). Four chains (three heated) were run for 2,000,000 generations with the default settings. Pilot runs using the more complex models (JC+I+G and JC+Cov) were examined for convergence in Tracer version 1.4 [22] and used to choose an appropriate burnin (sump and sumt burnin=5000; this equals 50,000 generations). Maximum parsimony (MP) analysis was conducted using PAUP* version 4.0b10 (with default settings except for HSEARCH NBEST=1).

For the model incorporating covarion evolution (JC+Cov), we used the covarion model of Tuffley and Steel [19]. Huelsenbeck [23] described an extension to this model with an underlying variable rates across sites (a rate for each site is first drawn from a gamma distribution) and an overlaying covarion process. Under this model a site can be variable, in which case its rate is taken from the gamma distribution, or invariable; an invariable site can become variable and vice versa. This model is implemented in a Bayesian framework in MrBayes. However, we encountered problems when using JC with variable rates across sites and covarion (JC+Hue). In many cases, the application of both these models to our data resulted in convergence on positive log likelihoods! Similar problems with MCMC using parameter rich models have been previously reported [24]. We reported these problems in April 2008 using the MrBayes bug report tool

(http://sourceforge.net/tracker/index.php?func=detail&aid=1945304&group_id=129302&atid=714418).

3.3.3 Reconstruction Accuracy

We evaluated the accuracy of the different analyses in reconstructing the tree T=((Group 1, Group 2),(Group 3, Group 4)) i.e. the inner-most edge splitting groups 1 and 2 from groups 3 and 4 (see Figure 3.1). The tree T is one of three possible trees splitting the four groups into two bipartitions (1+2 vs. 3+4, 1+3 vs. 2+4, and 1+4 vs. 2+3). For the Bayesian analyses, the support for each of the three possible trees was calculated as the number of datasets for which the tree had the highest frequency in the posterior distribution. For MP, the support for each of the three possible trees was calculated as the number of datasets for which the tree was inferred.

3.3.4 Model fit

There is no agreed-upon method for objective model selection in a Bayesian framework [25]. Therefore, we used several procedures to determine the best-fit model: (1) The Akaike Information Criterion (AIC; [26]) applied to the arithmetic mean of the estimated marginal likelihoods (as in [27]), (2) the AIC applied to the maximum

likelihood found for the cold chain (3) Bayesian Information Criterion (BIC; [28]) applied to the maximum likelihood found for the cold chain, and (4) Bayes factors (BF) applied to the harmonic mean of the estimated marginal likelihoods. The adequacy of each of the models was also evaluated using our own implementation of the method described by Bollback [15], which uses the posterior predictive distributions to account for uncertainty in the phylogeny and model parameters (the code is available from l.shavit@massey.ac.nz). This method assumes that an adequate model should perform well in predicting future observations. In absence of future observations (which is generally the case), predicted observations are simulated under the model in question by sampling from the joint posterior density of trees and parameters as approximated using MCMC. A test statistic is then used to evaluate the difference between the simulated and original data. This is a Bayesian analog of frequentist methods such as the classic parametric bootstrap [15]. We used the multinomial test statistic $T(X) = \left(\sum_{\xi \in S} N_\xi \ln\left(N_\xi\right) \right) - N \ln\left(N\right)$, where S is the set of (unique) possible site patterns, N is the number of sites, and N_ξ is the number of sites in which pattern ξ was observed. This is a general statistic which is used to test the overall predictive performance of the model rather than the performance of a specific aspect of the model. As in the phylogenetics analysis, the first 50,000 generations were discarded from the posterior distribution before conducting this analysis.

3.4 Results and Discussion

We evaluated tree reconstruction accuracies of Bayesian analyses using each of the five models (JC, JC+I, JC+G, JC+I+G, and JC+Cov), when applied to data where $Pvar^+$ = (0, 5, 10, 15, 20, 25, 30, 35, 40, 45, 50) percent of the invariable sites were reset to be variable in two events defined on the tree (Figure 3.1).

3.4.1 Tree Reconstruction Accuracy with Changing Proportions of Variable Sites – 4-taxa

Figure 3.3 shows the ability of the analyses to reconstruct the correct phylogeny for data that was simulated under JC without the covarion model, for the 4-taxon simulations. The only change in the evolutionary process is introduced (at the two events; see Figure 3.1) by an increased proportion of variable sites. In general, the higher the percentage of sites that become variable in the two events ($Pvar^+$) the less accurate the tree reconstruction is. None of the five models used for phylogenetic inference describe the data accurately (they do not account for the changing Pvar). Nevertheless, one might consider the JC+Cov model as the closest to the simulated data, as the changing proportions of variable sites are expected to produce covarion-like site patterns. However, the accuracy with which Bayesian analysis using this model (as well as JC) reconstructs the correct phylogeny is strongly impaired when $Pvar^+$ increases. For $Pvar^+ \geq 20\%$ the wrong tree (where the two non-sister lineages H and I, in which the change in Pvar occurred, are grouped together) is chosen most often. This may be, in part, due to the proportion of sites that are invariable across all taxa which is not accounted for by this model. For the JC+I model, the correct tree is chosen most often, although decreased accuracy is observed. The models allowing for variable rates across sites (JC+G and JC+I+G) are the most accurate in reconstructing the correct phylogeny for the parameters used in this simulation. Nevertheless, tree reconstruction under these models has been shown to be inconsistent when applied to other types of heterotachy [29, 30].

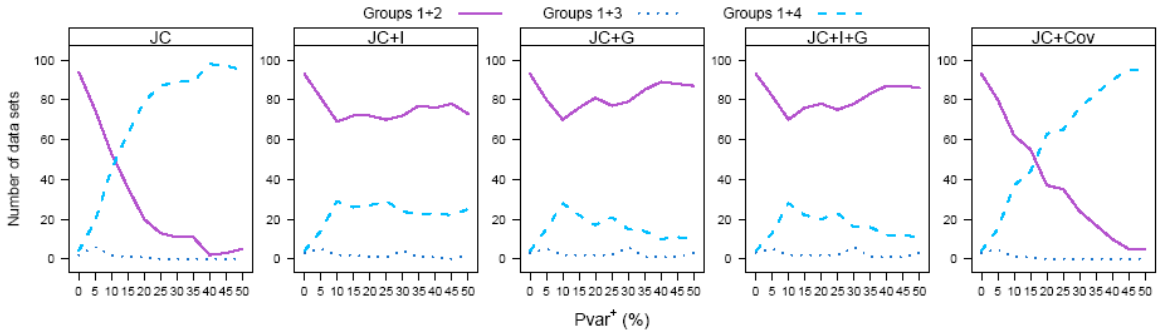


Figure 3.3 - Tree reconstruction accuracy for the 4-taxon simulations without the covarion model. Bayesian analysis was done using Jukes-Cantor (JC), JC with invariable sites (JC+I), JC with a gamma distribution (JC+G), JC with invariable sites and a gamma distribution (JC+I+G), and JC with the covarion model (JC+Cov). For each model, the sum of the proportional frequencies of each of the three possible splits of the groups (1+2 vs. 3+4, 1+3 vs. 2+4, and 1+4 vs. 2+3) is shown for an increasing Pvar$^+$ = (0, 5, 10, 15, 20, 25, 30, 35, 40, 45, 50) percent of the invariable sites that were reset to be variable in the two events.

3.4.2 Correlated vs. Uncorrelated Events

Next we tested the effect of the correlation between the two events. Correlated events, where the positions of sites that switch state are identical, might occur if a similar change in function (and therefore functional constraints) takes place in separate lineages. Conversely, uncorrelated events, where the positions of sites that switch state are independent, might occur when the change in constraints acting on the lineages is different. Tree reconstruction accuracies for the 4-taxon tree T_4 in the case of correlated events (Figure 3.3) were compared with the case of uncorrelated events (results not shown). We found that the effect of changing Pvar is much less pronounced in the case of uncorrelated events. In fact, the tree reconstruction accuracy of Bayesian analysis using any of the five models tested was higher than 86% for all values of Pvar$^+$. These results are expected, as the positions of sites that become variable at the events, in the two non-sister lineages (taxa H and I), are likely to be much less similar in this case (compared with the identical positions in the correlated case).

3.4.3 Adding the Covarion Model

Under the settings used in our simulations, having sites evolve under the covarion model raises the overall number of invariable sites (see Figure 3.2). This is done in a random manner (effectively reducing the correlation between the events) and therefore decreases the similarity between the positions of invariable sites in the two non-sister taxa H and I. This can be seen as an intermediate case between correlated and un-correlated events. We compared the tree reconstruction accuracies for data that was simulated with and without the covarion model (Figure 3.3 and Figure 3.4). The results show that when data is simulated without the covarion model (Figure 3.3) the effect of change in Pvar on phylogenetic inference is twice as strong as that when data is simulated with the covarion model (Figure 3.4). The inclusion of the covarion models delays, but does not change the nature of, the effect of changes in Pvar on tree reconstruction accuracy.

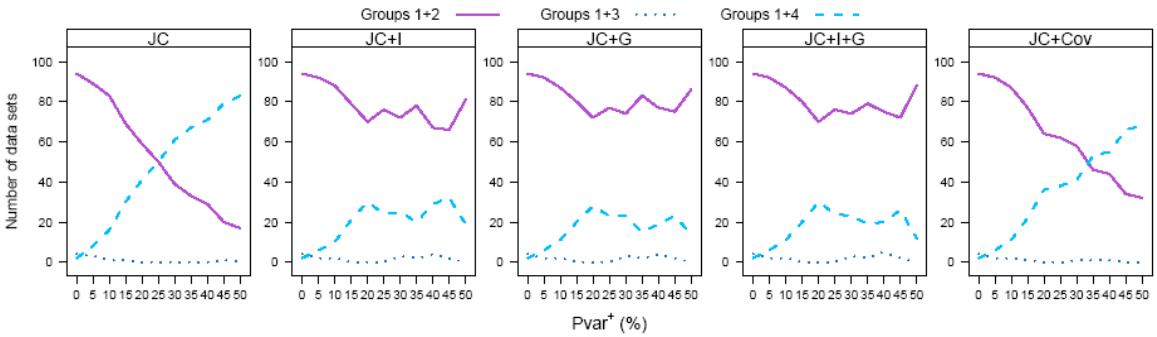


Figure 3.4 - Tree reconstruction accuracy for the 4-taxon simulations with the covarion model. Bayesian analysis was done using JC, JC+I, JC+G, JC+I+G, and JC+Cov. For each model, the sum of the proportional frequencies of each of the three possible splits of the groups (1+2 vs. 3+4, 1+3 vs. 2+4, and 1+4 vs. 2+3) is shown for an increasing $Pvar^{+}$ = (0, 5, 10, 15, 20, 25, 30, 35, 40, 45, 50).

3.4.4 Model Fit

Reconstructing trees under the best-fit model found using selection methods (e.g. as implemented in ModelTest [31] is a common procedure in phylogenetic inference.

However, model selection in a Bayesian framework is not straightforward. BF evaluate the evidence provided by the data in favor of one model over another [32]. Such pair-wise comparisons are useful, but model selection from a larger set of models is difficult and the results might depend on the order of pair-wise comparisons (the same problem is encountered when using likelihood ratio tests in a maximum likelihood framework [33]. In addition, the interpretation of BF is subjective. We therefore used the AIC and BIC, in addition to BF, to determine the best-fit model for each data set and compared their outcomes. The numbers of times with which each of the five models was found as the best-fit model using the AIC and BIC are shown in Figure 3.5.

For data simulated without the covarion model (Figure 3.5a), when $Pvar^+$ (the percentage of invariable sites that become variable in the two events) is zero or very small JC+I is chosen most frequently as the best-fit model. Indeed, for $Pvar^+=0$ this is the correct model. However, as $Pvar^+$ increases, the JC+Cov model is selected most often using both the AIC ($Pvar^+>15\%$) and BIC ($Pvar^+>25\%$). For data simulated with the covarion model (Figure 3.5b) when $Pvar^+$ is zero or very small JC+I is selected most frequently. For larger $Pvar^+$ using BIC ($Pvar^+>20\%$), JC+Cov and JC+G are alternatively chosen as the best fit model; using AIC, the JC+Cov and JC+G models are alternatively chosen as the best fit model for the middle range $Pvar^+$, and when $Pvar^+ \geq 35\%$ JC+Cov is most frequently chosen. The Bayes factor in favor of model 1 over model 0, B_{10}, was calculated for each dataset and each pair of models. The resulting BF were then interpreted according to the Kass and Raftery [32] version of the guidelines presented by Jeffreys [34]. The number of times a positive ($2\ln(B_{10})>2$) or strong ($2\ln(B_{10})>6$) support for favoring one model over another was summarized (online appendix 1). Overall, the larger $Pvar^+$ was, the higher the number of dataset for which the JC+Cov model was favored (this is congruent with the AIC and BIC results).

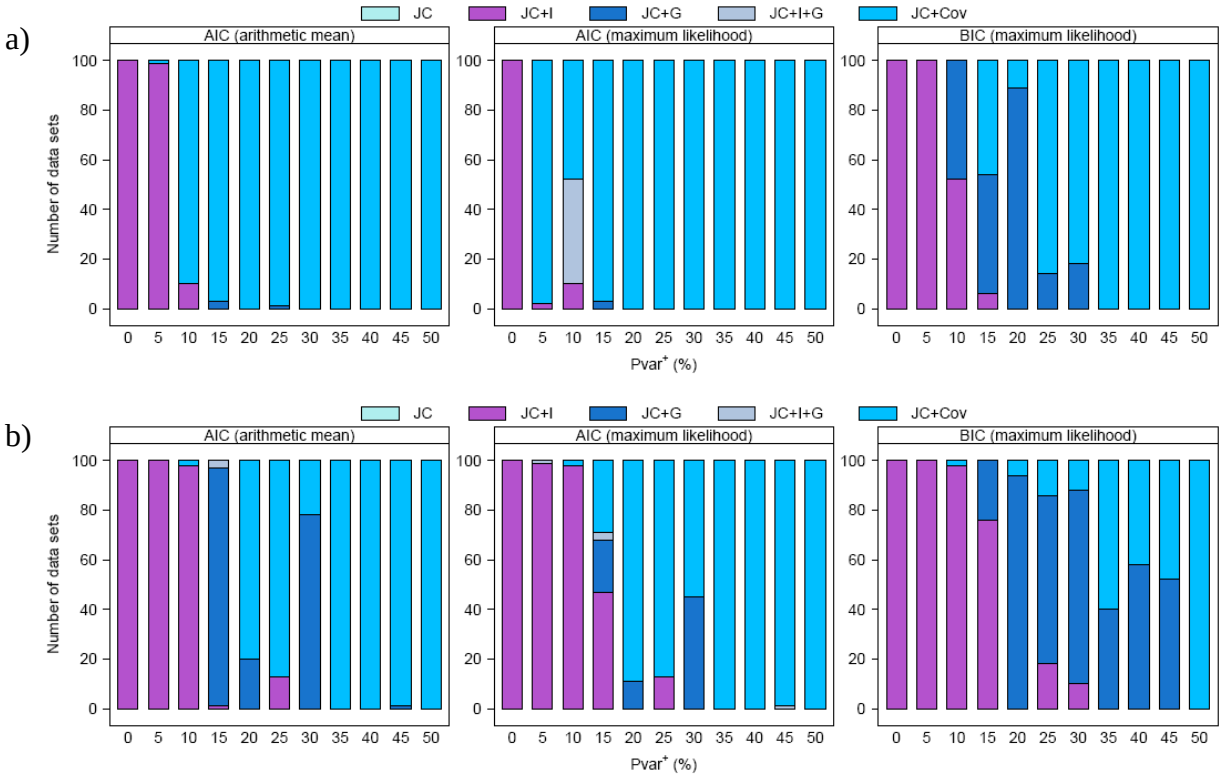


Figure 3.5 - Best-fit model for the 4-taxon simulations without (a) and with b) the covarion model. Comparison of the number of times each of the five models (JC, JC+I, JC+G, JC+I+G, and JC+Cov) was found to be the best-fit model using the Akaike Information Criterion (AIC) applied to both the arithmetic mean of the estimated marginal likelihoods and the maximum likelihood found for the cold chain, and the Bayesian Information Criterion (BIC) applied to the maximum likelihood found for the cold chain.

In our simulations, for $Pvar^{+} \geq 20\%$ with no covarion and $Pvar^{+} \geq 35\%$ when covarion was incorporated, using the best-fit model (JC+Cov) resulted in erroneous phylogenetic estimates more frequently than correct estimates. We then determined the adequacy of the best-fit model JC+Cov, as well as the other models, using the posterior predictive distributions (Figure 3.6; see methods for more detail). As the simulated change in Pvar increases, so does the number of datasets for which the JC+Cov model was rejected. Even when no change in Pvar was simulated ($Pvar^{+}=0$), these models were rejected for more than 29% of the datasets at the 1% level and 82% at the 5% level (not shown).

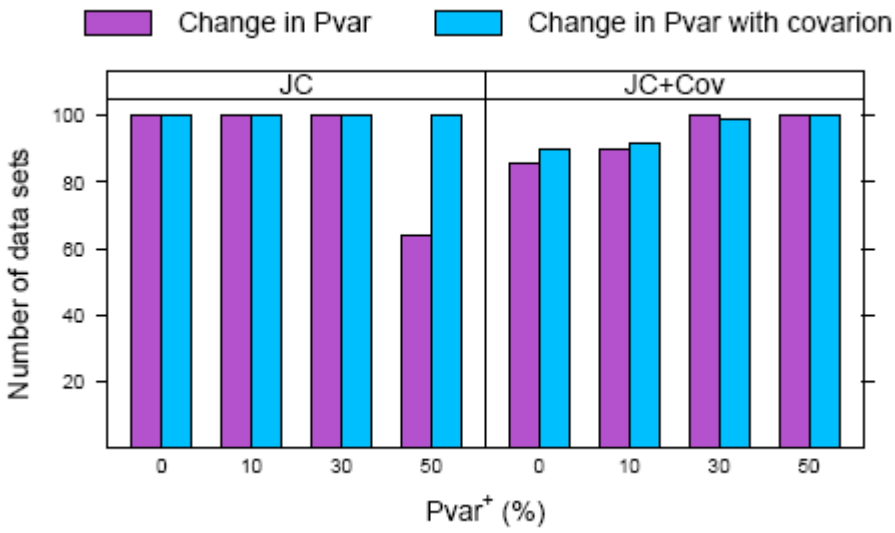


Figure 3.6 - Absolute model adequacy assessment for data simulated with and without the covarion model for an increasing Pvar^{+} = (0, 10, 30, 50). The number of times each model (JC, JC+Cov) was rejected at the 1% level. The JC+I, JC+G, and JC+I+G were never rejected.

These results, together with our tree reconstruction results in Figure 3.3 and Figure 3.4, suggest that the covarion model used (which assumes a constant number of variable sites) is inadequate at capturing change in proportions of variable sites. This simple covarion model is a priori disadvantaged in the case of our simulated data, as it does not account for the proportion of sites that is invariable throughout the tree. Unfortunately (and although not apparent), the JC+I+Cov model which is expected to fit our data relatively well is not implemented in MrBayes (in fact, any combination of I+Cov is not implemented. Several published papers, including our own, state that the model used was the I+Cov [17] (Chapter 2 in this thesis) or G+I+Cov [35, 36] but in practice the program ignores the I parameter only accounting for invariable sites under the covarion model).

When the covarion model was not used in simulations, the number of datasets for which the JC model was rejected decreased as the change in Pvar increased. When the covarion model was used, this trend was less pronounced. This might be predicted, considering that the JC model does not account for a constant proportion of invariable sites. The increased Pvar effectively decreases the number of invariable sites in the

dataset. In contrast, the addition of covarion sites effectively increases the number of invariable sites in the dataset. Notably, the three models that were found adequate for the data also performed well in tree reconstruction, whereas the two models that failed the absolute adequacy assessment all displayed lower tree reconstruction accuracy. The simple multinomial statistic used was able to identify model inadequacy, which was probably a result of these models inability to correctly account for the proportion of invariable sites.

3.4.5 Taxon Sampling

We investigated the effect of taxon sampling on the accuracy of the Bayesian phylogenetic inference by comparing the reconstruction accuracies of the tree T=((Group 1, Group 2),(Group 3, Group 4)) for the 4-, 8-, and 16-taxon simulations. The performance of all five models was evaluated for the 4- and 8-taxon simulation. For the 16-taxon simulations however only the best-fit model JC+Cov was evaluated. A comparison of the reconstruction accuracies using JC+Cov model is shown in Figure 3.7. Reconstruction accuracies using JC+Cov were similar (not shown). With the addition of taxa, the accuracy with which the correct split (Groups 1+2 vs. Groups 3+4) is found increases significantly. For the 8-taxon simulations, the correct split is found most often using any of the six models (results not shown). These findings are in agreement with earlier observations [12, 29, 37].

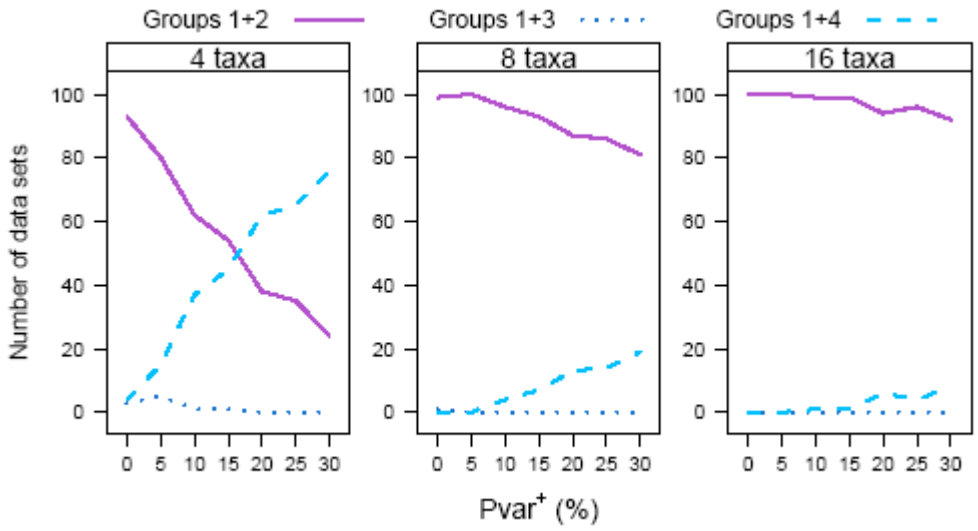


Figure 3.7 - The effect of taxon sampling on reconstruction accuracy of the main split of the tree T (Groups 1+2 vs. Groups 3+4). The reconstruction accuracy for the 4-, 8-, and 16-taxon simulations using the JC+Cov model is shown for an increasing Pvar$^+$ = (0, 5, 10, 15, 20, 25, 30).

In order to distinguish between improved accuracy due to increased taxon sampling in general versus more extensive sampling of taxa subsequent to the two events, we evaluated the accuracy of phylogenetic inference using the JC+Cov model when applied to two different 6-taxon trees. Tree T_{6a} contains two taxa under each of the two events (Groups 2 and 3) and one taxon under each of the other two lineages (Groups 1 and 4), whereas tree T_{6b} contains only one taxon under each of the two events and two taxa under each of the other two lineages (Figure 3.1). We found (Figure 3.8Figure 3.) that increased taxon sampling in lineages that did not undergo change in Pvar (T_{6b}) does not improve the reconstruction accuracy of the main split of the tree T (in comparison to tree reconstruction accuracy for the 4-taxon simulations), whereas increased taxon sampling in the lineages under the two events (T_{6a}) improves the tree reconstruction accuracy and delays the accuracy hindering effect of change in Pvar.

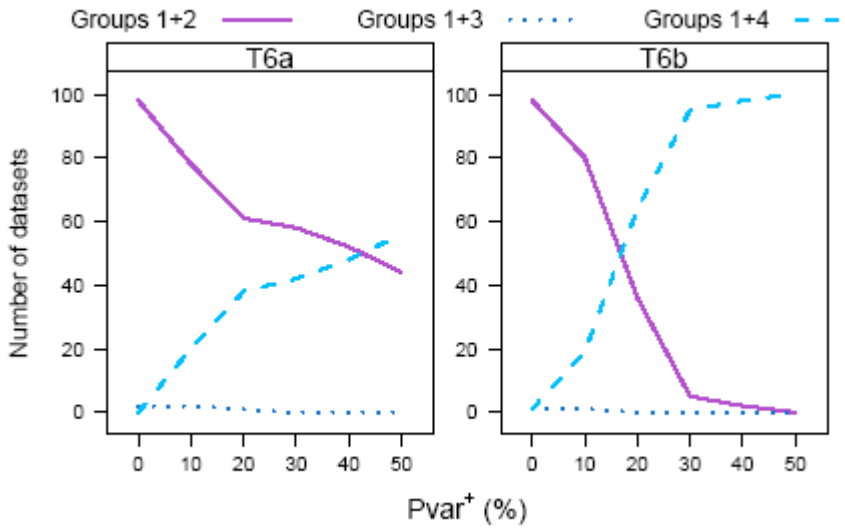


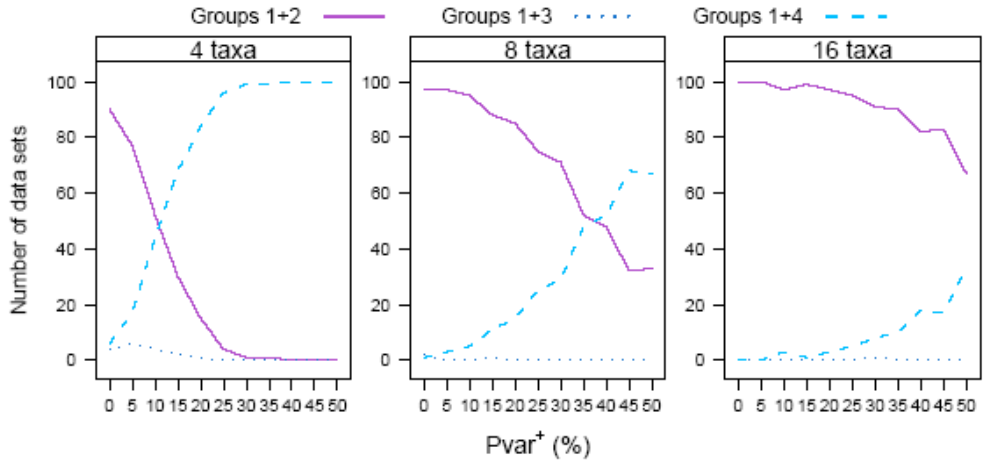
Figure 3.8 - Comparison of reconstruction accuracy of the main split of the tree T (Groups 1+2 vs. Groups 3+4) for general increased taxon sampling versus increased taxon sampling under the two events. The tree reconstruction accuracy for the data simulated under T_{6a}=((A,(E,H)),((I,L),P)) and T_{6b}=(((A,D),H),(I,(M,P))) using the JC+Cov model is shown for an increasing $Pvar^{+}$ = (0, 10, 20, 30, 40, 50).

3.4.6 Maximum Parsimony vs. Bayesian Analysis

Kolaczkowski and Thornton [30] reignited a two-decades long debate when they claimed that Maximum Parsimony (MP) performs better than Maximum Likelihood (ML) and Bayesian analysis for a range of parameters. The authors' conclusion was based on a very specific case of heterotachy (convergent change in overall rates in non-sister lineages) with a specific combination of parameters and tree topology. Several contradicting results were later published [38, 39, 40] and the biological realism of the original work has been questioned [41] (but see [42]). Kolaczkowski and Thornton further declared that MP is unaffected by heterotachy [30, 42]. However, Philippe et al. [40] later showed that when the level of rate variation across lineages (level of convergent change in overall rates in non-sister lineages) increases, MP accuracy can either decrease or increase depending on the relative branch lengths. To shed further light on this debate, we present a comparison of the accuracy of MP with that of Bayesian analysis using JC+Cov in reconstructing the correct tree T (see Figure 3.1).

Phylogenetic inference using MP was applied to the 4-, 8-, and 16-taxa simulations. The accuracy with which MP reconstructs the correct phylogeny is greatly hindered by the increased $Pvar^+$ (Figure 3.9a). The increase in taxon sampling does improve MP accuracy, however, the Bayesian analyses (Figure 3.7) were found to be more accurate than MP and less affected by the increased $Pvar^+$. We also tested the ability of MP to reconstruct the 4-taxon tree (T_4) in the case of uncorrelated events. MP is clearly affected by the increased $Pvar^+$ (Figure 3.9b) with the wrong tree where the two non-sister lineages are grouped together reconstructed most frequently when $Pvar^+$>0.35.

a)



b)

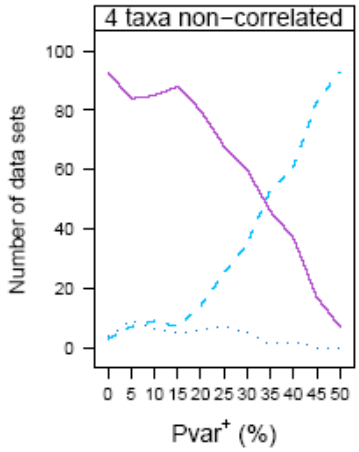


Figure 3.9 - Tree reconstruction accuracy using maximum parsimony (MP). a) The effect of taxon sampling on reconstruction accuracy of the main split of the tree T (Groups 1+2 vs. Groups 3+4). b) tree estimation for the 4-taxon simulations with uncorrelated events (the positions of sites that switch state are independent). The tree reconstruction accuracy is shown for an increasing $Pvar^+$ = (0, 5, 10, 15, 20, 25, 30, 35, 40, 45, 50).

3.5 **Conclusions**

Change in the proportions of variable sites causes a model misspecification that can mislead phylogenetic methods. We found that a simple covarion model is inadequate at capturing such changes. A model combining a proportion of sites that are invariable across the tree and covarion evolution is not currently implemented in MrBayes. Although this model does not account for changes in the proportion of variable sites, it is expected to fit such data relatively well. Testing the ability of this model to reconstruct trees from simulated data containing change in the proportion of variable sites is important for our understanding of the effects of model misspecification of this kind.

Our results show that the use of the best-fit model, chosen by a relative criterion, does not guarantee correct tree reconstruction. In fact, the best-fit model for our data performed poorly, while other models performed better, and absolute model-fit assessments confirmed that this best-fit model is inadequate for our data. Although none of the tested models accounts for changes in Pvar, some of the models could not be rejected by the absolute model-fit assessment. Importantly, these models were more accurate in tree reconstruction. Further work to test the performance of relative and absolute model-fit tests for a large number of trees and a wide range of parameters is needed before a general conclusion can be drawn. We therefore recommend the use of absolute model-adequacy tests [14, 15], alongside relative-fit tests, as an integral part of phylogenetic analysis.

We found that taxon sampling has a strong effect on tree reconstruction accuracy. In particular, greater taxon sampling under the events in which a change in Pvar occurred resulted in improved accuracy. Our results imply that more accurate phylogenetic inference can be achieved by inclusion of larger numbers of taxa from lineages for which prior knowledge suggests that a change in the evolutionary process occurred.

In contrast with the reports of Kolaczkowski and Thornton [30, 42], we establish that the accuracy of MP can be adversely affected by heterotachy. Increase in taxon sampling did improve the accuracy of MP, yet it was still the least accurate in tree reconstruction.

Currently implemented phylogenetic models do not account for changes in the proportions of variable sites. This model misspecification can result in erroneous tree reconstruction. However, the accuracies of tree estimation using different models vary; and although not accounting for heterotachy, a model can sometimes be adequate for heterotachous data. An absolute goodness of fit test is useful in evaluating model adequacy and can help differentiate cases in which tree reconstruction is expected to be accurate, from cases in which the model is inadequate and its use is likely to result in incorrect tree estimation. Branch-length mixture models that aim to account for heterotachy [10, 43] exist. Testing the accuracy of such models to the data containing changes in Pvar (such as that simulated here) would be an interesting extension of the present study.

3.6 **Acknowledgements**

We thank Pete Lockhart for helpful discussions. We also thank Jack Sullivan, Olivier Gascuel, and three anonymous referees for their constructive comments that helped us improve this manuscript. This work was financially supported by the New Zealand Marsden fund (05-MAU-033 to B.R.H).

3.7 References

1. Philippe H and Lopez P. 2001. On the conservation of protein sequences in evolution. Trends Biochem Sci 26:414-416.
2. Fitch WM and Markowitz E. 1970. An Improved Method for Determining Codon Variability in a Gene and Its Application to Rate of Fixation of Mutations in Evolution. Biochem Genet 4:579-593.
3. Lopez P, Casane D and Philippe H. 2002. Heterotachy, an important process of protein evolution. Mol Biol Evol 19:1-7.
4. Ane C, Burleigh JG, McMahon MM and Sanderson MJ. 2005. Covarion structure in plastid genome evolution: A new statistical test. Mol Biol Evol 22:914-924.
5. Lockhart P, Novis P, Milligan BG, Riden J, Rambaut A and Larkum T. 2006. Heterotachy and tree building: A case study with plastids and eubacteria. Mol Biol Evol 23:40-45.
6. Gruenheit N, Lockhart PJ, Steel M and Martin W. 2008. Difficulties in testing for covarion-like properties of sequences under the confounding influence of changing proportions of variable sites. Mol Biol Evol 25:1512-1520.
7. Inagaki Y, Susko E, Fast NM and Roger AJ. 2004. Covarion shifts cause a long-branch attraction artifact that unites microsporidia and archaebacteria in EF-1 alpha phylogenies. Mol Biol Evol 21:1340-1349.
8. Germot A and Philippe H. 1999. Critical analysis of eukaryotic phylogeny: a case study based on the HSP70 family. J Eukaryot Microbiol 46:116-124.
9. Lockhart PJ and Steel MA. 2005. A Tale of Two Processes. Syst Biol 54:948-951.
10. Kolaczkowski B and Thornton JW. 2008. A mixed branch length model of heterotachy improves phylogenetic accuracy. Mol Biol Evol.

11. Holland BR, Penny D and Hendy MD. 2003. Outgroup misplacement and phylogenetic inaccuracy under a molecular clock - A simulation study. Syst Biol 52:229-238.

12. Shavit L, Penny D, Hendy MD and Holland BR. 2007. The Problem of Rooting Rapid Radiations. Mol Biol Evol 24:2400-2411.

13. Minin V, Abdo Z, Joyce P and Sullivan J. 2003. Performance-Based Selection of Likelihood Models for Phylogeny Estimation. Syst Biol 52:674-683.

14. Goldman N. 1993. Statistical Tests of Models of DNA Substitution. J Mol Evol 36:182-198.

15. Bollback JP. 2002. Bayesian model adequacy and choice in phylogenetics. Mol Biol Evol 19:1171-1180.

16. Posada D and Buckley TR. 2004. Model selection and model averaging in phylogenetics: advantages of akaike information criterion and bayesian approaches over likelihood ratio tests. Syst Biol 53:793-808.

17. Shavit Grievink L, Penny D, Hendy MD and Holland BR. 2008. LineageSpecificSeqgen: generating sequence data with lineage-specific variation in the proportion of variable sites. BMC Evol Biol 8:317.

18. Rambaut A and Grassly NC. 1997. Seq-Gen: An application for the Monte Carlo simulation of DNA sequence evolution along phylogenetic frees. Computer Applications in the Biosciences 13:235-238.

19. Tuffley C and Steel M. 1998. Modeling the covarion hypothesis of nucleotide substitution. Math Biosci 147:63-91.

20. Jukes TH and Cantor CR. 1969. Evolution of protein sequences. Pp. 21-123 in Munro HN, ed. Mammalian protein metabolism. Academic Press, New York.

21. Ronquist F and Huelsenbeck JP. 2003. MrBayes 3: Bayesian phylogenetic inference under mixed models. Bioinformatics 19:1572-1574.

22. Rambaut A and Drummond A. 2007. Tracer v1.4. Available from http://beast.bio.ed.ac.uk/Tracer.

23. Huelsenbeck JP. 2002. Testing a covariotide model of DNA substitution. Mol Biol Evol 19:698-707.

24. Smedmark JEE, Swenson U and Anderberg AA. 2006. Accounting for variation of substitution rates through time in Bayesian phylogeny reconstruction of Sapotoideae (Sapotaceae). Mol Phylogenet Evol 39:706-721.

25. Huelsenbeck JP, Larget B, Miller RE and Ronquist F. 2002. Potential Applications and Pitfalls of Bayesian Inference of Phylogeny. Syst Biol 51:673-688.

26. Akaike H. 1974. A new look at the statistical model identification. Automatic Control, IEEE Transactions on 19:716-723.

27. Strugnell J, Norman M, Jackson J, Drummond AJ and Cooper A. 2005. Molecular phylogeny of coleoid cephalopods (Mollusca: Cephalopoda) using a multigene approach; the effect of data partitioning on resolving phylogenies in a Bayesian framework. Mol Phylogenet Evol 37:426-441.

28. Schwarz G. 1978. Estimating the Dimension of a Model. The Annals of Statistics 6:461-464.

29. Ruano-Rubio V and Fares MA. 2007. Artifactual phylogenies caused by correlated distribution of substitution rates among sites and lineages: the good, the bad, and the ugly. Syst Biol 56:68-82.

30. Kolaczkowski B and Thornton JW. 2004. Performance of maximum parsimony and likelihood phylogenetics when evolution is heterogeneous. Nature 431:980-984.

31. Posada D and Crandall KA. 1998. MODELTEST: testing the model of DNA substitution. Bioinformatics 14:817-818.

32. Kass RE and Raftery AE. 1995. Bayes Factors. Journal of the American Statistical Association 90:773-795.

33. Sullivan J and Joyce P. 2005. Model selection in phylogenetics. Annu Rev Ecol Evol Syst 36:445-466.

34. Jeffreys H. 1961. Theory of probability, 3rd edition. Oxford University Press, Oxford, UK.

35. Hampl V, Vrlik M, Cepicka I, Pecka Z, Kulda J and Tachezy J. 2006. Affiliation of Cochlosoma to trichomonads confirmed by phylogenetic analysis of the small-subunit rRNA gene and a new family concept of the order Trichomonadida. Int J Syst Evol Microbiol 56:305-312.

36. Gittenberger E, Piel WH and Groenenberg DSJ. 2004. The Pleistocene glaciations and the evolutionary history of the polytypic snail species Arianta arbustorum (Gastropoda, Pulmonata, Helicidae). Mol Phylogenet Evol 30:64-73.

37. Heath TA, Zwickl DJ, Kim J and Hillis DM. 2008. Taxon sampling affects inferences of macroevolutionary processes from phylogenetic trees. Syst Biol 57:160-166.

38. Spencer M, Susko E and Roger AJ. 2005. Likelihood, parsimony, and heterogeneous evolution. Mol Biol Evol 22:1161-1164.

39. Gadagkar SR and Kumar S. 2005. Maximum likelihood outperforms maximum parsimony even when evolutionary rates are heterotachous. Mol Biol Evol 22:2139-2141.

40. Philippe H, Zhou Y, Brinkmann H, Rodrigue N and Delsuc F. 2005. Heterotachy and long-branch attraction in phylogenetics. BMC Evol Biol 5.

41. Steel M. 2005. Should phylogenetic models be trying to 'fit an elephant? Trends Genet 21:307-309.

42. Thornton JW and Kolaczkowski B. 2005. No magic pill for phylogenetic error. Trends Genet 21:310-311.

43. Pagel M and Meade A. 2008. Modelling heterotachy in phylogenetic inference by reversible-jump Markov chain Monte Carlo. Philos Trans R Soc Lond B Biol Sci 363:3955-3964.

Chapter 4

Change in evolutionary constraints and the long-branch attraction artifact

Manuscript in preperation; in collaboration with David Bryant, Peter Lockhart, and William Martin.

4.1 Abstract

Long-branch attraction is a well known phenomenon in molecular phylogenetic analysis. It is thought to affect tree reconstruction of many groups. Long branches (which cannot be attributed to time) are generally assumed to be fast evolving lineages. Therefore, much effort has been put towards accounting for differences in overall evolutionary rates across sites and across lineages. Focusing on the case of Microsporidia, a parasite whose basal positioning to Eukaryotes was suggested to be a result of long-branch attraction artifact, we address the problem of long-branch attraction from a novel perspective. Using a non-parametric bootstrap test, we identify changes in the evolutionary process manifested in the types of substitutions that take place along the Microsporidia branch. We found that long-branch estimates and basal positioning of Microsporidia both correlate with increased proportions of radical substitutions in the microsporidian lineage. We show that such changes in the substitution process can lead to erroneous inference of long branches. We conclude that erroneous long-branch inferences are likely to be the result of changes in the substitution process, which are not accounted for by commonly used phylogenetic models.

4.2 **Introduction**

Despite the immense progress that has been made in molecular phylogenetics over the last few decades, many phylogenies remain unresolved. A major obstacle in reconstructing reliable phylogenies from molecular data is long-branch attraction [1, 2, 3], a common systematic error where two non-adjacent long branches are mistakenly grouped together. Long-branch attraction (LBA) has been suggested to affect tree reconstruction of many groups, including angiosperms [4], birds [5], bees [6], fishes [7], and mammals [8]. It was also suggested to be the reason why some early phylogenetic analyses failed to recover unquestioned relationships such as the monophyly of rodents [9]. The long-branch attraction artifact is of particular concern when a distant outgroup is used in tree reconstruction for rooting and molecular dating. In this case, the long-branch lineages of the ingroup are 'attracted' by the outgroup lineage causing an artificial early emergence of the long-branch ingroup lineages [10, 11, 12 (Appendix in this thesis)].

How should these ingroup long branches be interpreted? Most researchers consider long branches an indication of fast rates of substitution and the long-branch lineages as "fast evolving lineages" [13, 14]. In contrast, the study presented here aims to test the hypothesis that long branches, rather than being an indication of fast rates, are the result of reduced functional and structural constraints on the sequence. These constraints, as initially suggested by Dickerson [15], can vary with change in 3D structure over time and along the sequence; they determine, at any given time, what (proportion and positions) of sites are free to vary and the types of substitutions that can occur at any site.

We focus on the intriguing case of Microsporidia, highly reduced intracellular parasites of eukaryotes, which are known to cause diseases in many animals including humans. Microsporidia lack several eukaryotic features such as peroxisomes, flagella, and Golgi membranes but they possess remnant mitochondria (mitosomes) [16]. Until recently they were considered to be primitive and early branching eukaryotes [17, 18, 19, 20],

but this view changed as analyses of many other protein-coding genes have positioned Microsporidia as sister to, or included within, Fungi [19, 20, 21, 22, 23, 24]. The within-eukaryotic crown position is now widely accepted to be correct; and the basal placement of Microsporidia in earlier analyses is considered to be the result of a LBA artifact, where the long Microsporidia branch is attracted to the long branches leading to the Archaea outgroup [25]. Inagaki et al. [25] have suggested that this long-branch attraction is caused by site-specific rate variation through time. Brinkmann et al. [35] assembled a relatively large dataset containing 44 species and 133 genes; the Microsporidian *Encephalitozoon cuniculi* is included in 122 of these genes. Analysis of this dataset [26] has revealed changes in the base frequencies along the sequences. Rate variations through time and changes in base frequencies could both be explained by the process of changing constraints on the possible types of amino acids substitutions.

We speculate that the branch length of Microsporidia is over-estimated due to relaxation of constraints in the Microsporidia lineage, causing an increased degree of freedom for the possible types of amino-acid substitution at a site. This is manifested in a higher proportion of radical substitutions (Prad), i.e. substitutions between amino acids with different chemical properties. In order to test whether long branches can be attributed to higher Prad we develop a non-parametric bootstrap test to identify changes in the relative rates of amino-acid substitutions.

4.3 Materials and Methods

4.3.1 Identifying variations in the types of amino acids substitutions

We introduce a non-parametric bootstrap test which is based on the estimated normalized instantaneous rate matrices for each pair of taxa. These matrices can be calculated (from an alignment) in different ways, one of which is described in supplementary I. The matrices are then normalized in order to eliminate the variation in the overall molecular substitution rates. The test considers one taxon at a time; this enables us to identify the taxa in which changes have occurred. First, the average of all pairwise instantaneous rate matrices is calculated. Then, for each taxon, the average of pairwise instantaneous rate matrices that do not involve that taxon is calculated. The matrix describing the difference between these two average matrices is used to calculate a distance measure. This distance measure is calculated for the observed dataset and each of the bootstrap alignments. Informally, the further this distance measure deviates from zero the more different the process of substitution is for the taxon under consideration. To evaluate the significance of this distance we utilize the bootstrap technique.

The null hypothesis is that the taxon under examination has evolved under the same evolutionary constraints (the same substitution rate matrix) as the rest of the taxa in the dataset. If the null hypothesis is true than the distance measure for the observed data should not differ significantly from zero. The distribution of the distance measures of the bootstrap alignments, around the estimated distance measure (for the observed data), is approximately the same as the variance of the estimated distance measure and the actual distance for the observed data. We can therefore compare the estimated distance measure for the observed data (minus the actual distance measure, which is assumed to be zero) to this distribution. If the estimated distance measure for the observed data falls outside of the distribution then the null hypothesis can be rejected, and the taxon is said to display different types of amino-acid substitutions than the rest of the taxa in the data set. The power of the test can be increased by clustering the amino acids according to

their chemical properties (for example, using the Dayhoff classification as is done later in this chapter), thus reducing the size of the matrices and the number of empty or near-empty cells. Though not developed here, this test can also be applied to nucleotide instantaneous rate matrices.

The non-parametric bootstrap test can be described more formally as follows.

a. For each pair of taxa, *x* and *y*, estimate $Q^{x,y}$ and $Q^{y,x}$, where $Q^{i,j}$ is the *Q* matrix from taxon *i* to taxon *j* (see supplementary I)*;*

b. Calculate $\overline{Q}(\mathrm{X})$ and $Q_z(\mathrm{X})$, where X is the observed alignment, and

$$\overline{Q}(\mathrm{X}) = \frac{\sum_{x,y\in \mathrm{X}} \left(Q^{x,y} + Q^{y,x}\right)}{N(N-1)}$$

and

$$Q_z(\mathrm{X}) = \frac{\sum_{x,y\in \mathrm{X}} \left(Q^{x,y} + Q^{y,x}\right)}{(N-1)(N-2)}$$, where *N* is the number of taxa, *z* is the taxon under examination and $x, y \neq z$;

c. Calculate the difference measure $d(\mathrm{X})$, where X is the observed alignment and $d(\mathrm{X}) = \overline{Q}(\mathrm{X}) - Q_z(\mathrm{X})$.

d. Compute the test statistic $t = \left\| d(\mathrm{X}) - 0 \right\|_F$, where $\| \, . \, \|_F$ denotes the sum of squares of matrix entries.

e. Compute *n* non-parametric bootstrap alignments (resampling with replacement).

f. For each bootstrap alignment X_i, $1 \le i \le n$, calculate $d_i = \left\| d(\mathrm{X}_i) - d(\mathrm{X}) \right\|_F$. This results in a distribution d_i of values.

g. Compare t to the distribution of d_i values.

The distribution of $d(\mathrm{X}_i)$ around $d(\mathrm{X})$ is approximately the same as the distribution of $d(\mathrm{X})$ around the true distance between the *Q* matrices $d(true)$. If there is no significant difference between the *Q* matrices, then $d(true)$ should be close to 0 and so t should fall within the distribution of d_i values. If t falls outside or in the 5% tail of the distribution then we can reject the null hypothesis that taxon *z* has the same substitution process as the other taxa in the dataset.

For each taxon *z*, the proportion of radical substitutions (Prad) was calculated as follows:

$\mathrm{P}rad_z(\mathrm{X}) = \frac{\sum_{l,k} Q_z(\mathrm{X})_{l,k}}{\sum_{r,s} Q_z(\mathrm{X})_{r,s}}$, where l and r are row indices, and k and s are column indices of the rate matrix $Q_z(\mathrm{X})$; $l \neq k$, $r \neq s$ and the amino acids corresponding to indices *l* and *k* belong to different Dayhoff classes.

4.3.2 Simulations

4.3.2.1 Test validation

To validate our test we conducted simulations where a known change in process was introduced. Sequences were generated using our own R script, which makes use of the simSeq method in the R package Phangorn [27]. One hundred datasets of 10,000 amino acids were generated along the 8 taxon tree shown in Figure 4.1, with branch lengths interpreted as the expected number of substitutions per site. At the root the substitution rate matrix was set to be either the Dayhoff rate matrix, or one of five matrices each of which defines a different proportion *w* of within-Dayhoff-classes substitutions where *w* = (14.73 [equal relative rates], 20, 40, 60, 80) percent. The overall rate was assumed to be 1. The relative rate of each of the 28 types of within Dayhoff-classes substitutions

was calculated as $w/28$ and the relative rate of each of the 162 between Dayhoff-classes substitutions was calculated as $(1-w)/162$. A change in the substitution process (the relative rates within the rate matrix) was introduced in taxon T4 (Figure 4.1), which was (in all cases) simulated under equal relative rates. This simulates a relaxation in evolutionary constraints on the branch leading to taxon T4.

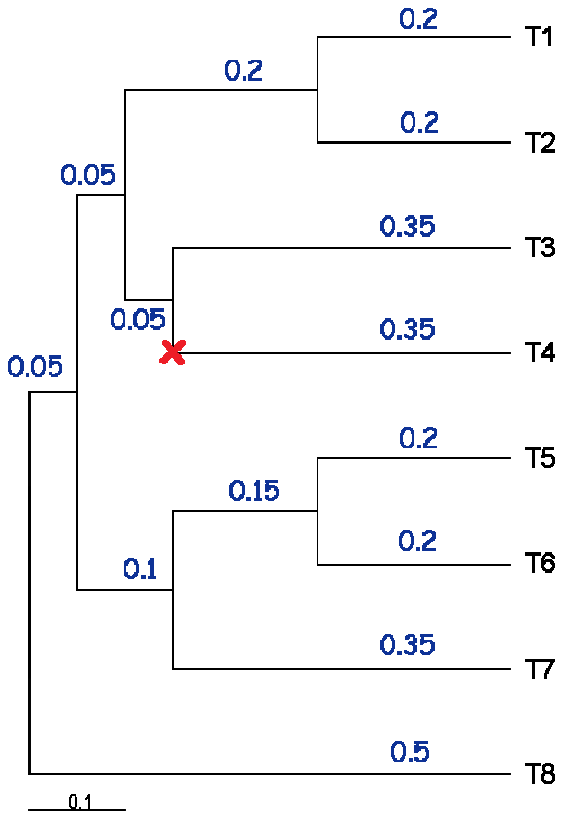


Figure 4.1 - The 8-taxon tree used for simulations. **A change in the substitution process (marked with X) was introduced in taxon T4.**

4.3.2.2 The effect of change in process on branch length

To demonstrate the effect of change in evolutionary process on tree reconstruction and branch estimation we simulated one hundred datasets, of 10,000 amino acids, under two different matrices (Dayhoff and equal relative rates) on the unrooted tree shown in Figure 4.2. We then reconstructed the trees using a (single) Dayhoff substitution matrix

and the average branch lengths (for the 100 trees) were compared with those of the tree used for simulation.

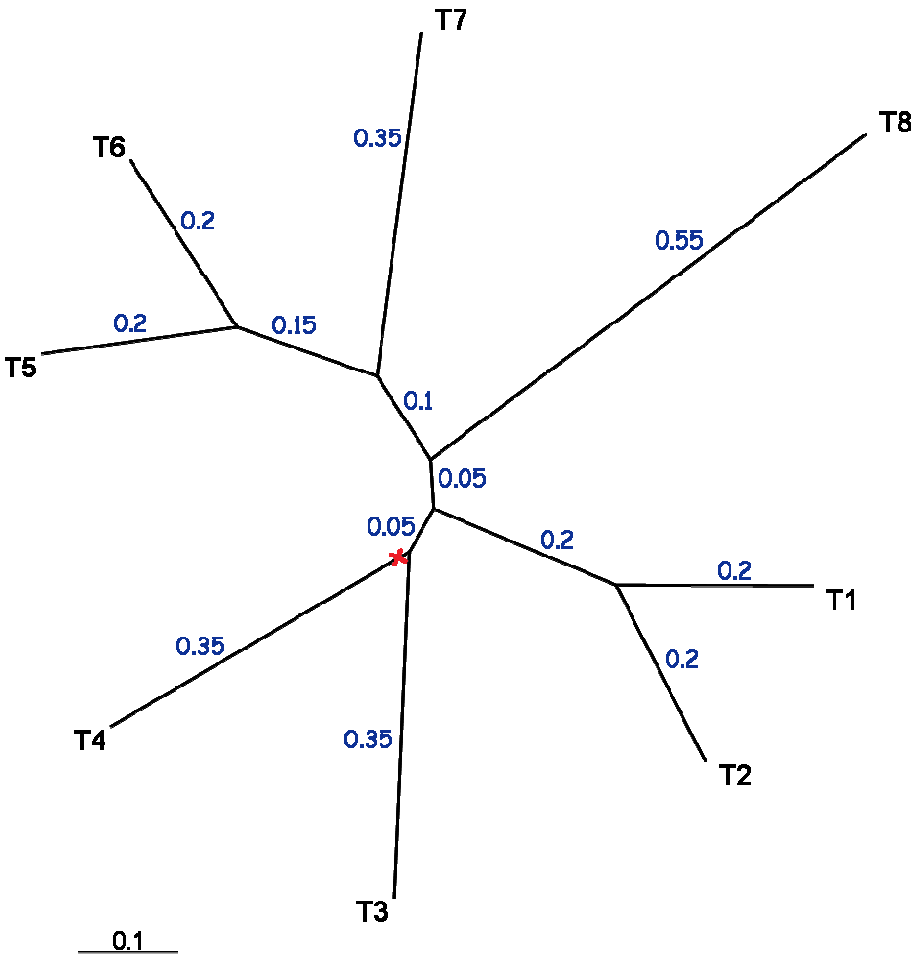


Figure 4.2 – The unrooted tree on which the sequences were simulated. The Dayhoff substitution matrix was used to simulated on all branches except for the branch of taxon T4 for which a matrix with equal relative rates was used.

4.3.3 Microsporidia data set

Brinkmann et al. [14] assembled a dataset containing 44 species and 133 genes with sequence lengths summing to 24,294 amino acid positions. We used 122 of these alignments in which the Microsporidian *Encephalitozoon cuniculi* is included. Taxa with more than 25% of the sites missing or unknown were removed from these alignments, as well as all sites with missing data.

4.3.4 Phylogenetic Reconstruction

Trees were estimated under the maximum likelihood criterion using the program PhyML (version 2.4). For the Microsporidia data the JTT+I+G substitution model was used. For the simulated data the Dayhoff+I+G substitution model was used.

4.4 Results and Discussion

4.4.1 Test validation

Simulated sequences were used to validate the level and power of the non-parametric bootstrap test. Figure 4.3 shows that the test can identify the introduced change in the substitution rate matrix of taxon T4. As the proportion of within-Dayhoff-classes substitutions, *w*, increases (i.e. the constraints on all other taxa are tightened), the equal relative rates of taxon T4 deviate more and more from the process in the rest of the taxa. Indeed the test detects more change in the substitution process in taxon T4 as the *w* in the other taxa increases (Figure 4.3).

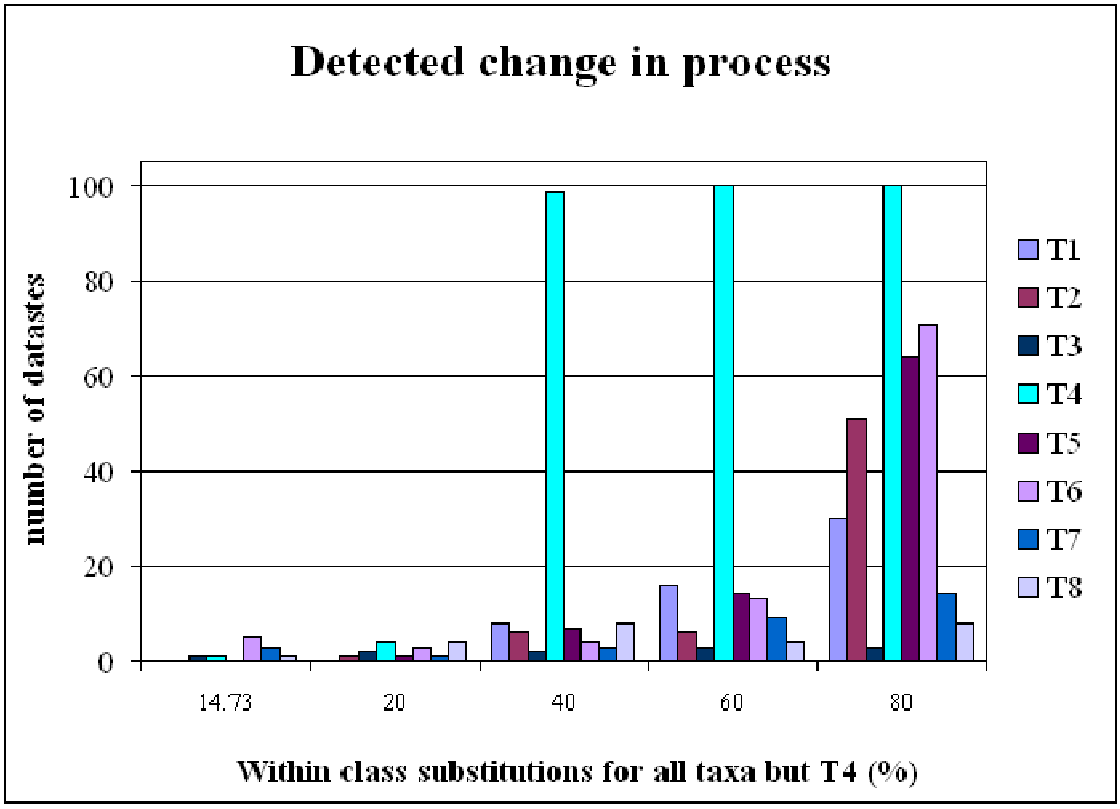


Figure 4.3 – The number of datasets in which a change in the substitution process was detected, for each taxon, as the proportion of within-class substitutions in all taxa but taxon T4 increases.

When all taxa evolve under equal relative rates (*w* = 14.73%), very little change is detected, i.e. there is a low number of false positives. The test is most accurate for the

moderate changes in process (*w*=40%, 60%). For *w* = 80%, the process under relaxed constraints in taxon T4 is extremely different from the process under very strong constraints in the rest of the taxa. Such an extreme difference causes a large number of false positives and change is detected in more taxa than just T4. This is due to the nature of the test; for any taxon apart for taxon T4, as all matrices involving the taxon under examination (*z*) are excluded to create $Q_z(\mathrm{X})$, one matrix is the comparison of that taxon with T4. As T4 is very different from the other taxa, removing this matrix will significantly change the average and will lead to an erroneous inference of change in the examined taxon. For the simulated datasets above, taxon T3 is relatively close to taxon T4 and so the inferred rate matrix for this pair of taxa does not affect the average matrix as in the case of other taxa and no change is detected for T3. Therefore the results of this test should be interpreted with caution, and when a change in process is detected in a large number of taxa we suggest removing the taxon (or group of taxa) for which prior knowledge (such as known change in functionality) suggest that a change in the substitution process has occurred and then applying the test to the rest of the data.

4.4.2 Test case: Microsporidia

The non-parametric bootstrap test using the full matrices (20*20) is not powerful enough for the Microsporidia dataset. Therefore, the matrices were reduced to 6*6 matrices (for the 6 Dayhoff classes). We detected a significant change in the substitution processes in the Microsporidia lineage in 85 of the 122 genes (see supplementary II). For each of these 85 genes, we also looked at the nature of the identified change. In particular, we have estimated Prad for each taxon and compared it to that of the other taxa (radical substitutions were defined to be substitutions between different Dayhoff classes). This was done by calculating Prad within the averaged rate matrix of all pairwise comparisons involving the taxon under consideration, and comparing it to Prad within the averaged matrix of all pairwise comparisons (i.e. $\overline{Q}(\mathrm{X})$).

An increase in the Prad was found in 80 out of the 85 genes in which a significant change was detected. In the other 5 genes slightly lower Prads were found for

Microsporidia. A higher Prad was detected in 59 out of the 66 genes in which Microsporidia is the longest terminal branch (Figure 4.4) and in 60 out of 72 genes in which Microsporidia is basal (Figure 4.5). For the 6 genes in which Microsporidia groups with Fungi, the test only detects one case in which Microsporidia has an increased proportion of radical substitutions. The Microsporidia branch for that gene is extremely long (Figure 4.6) explaining the identified change in process.

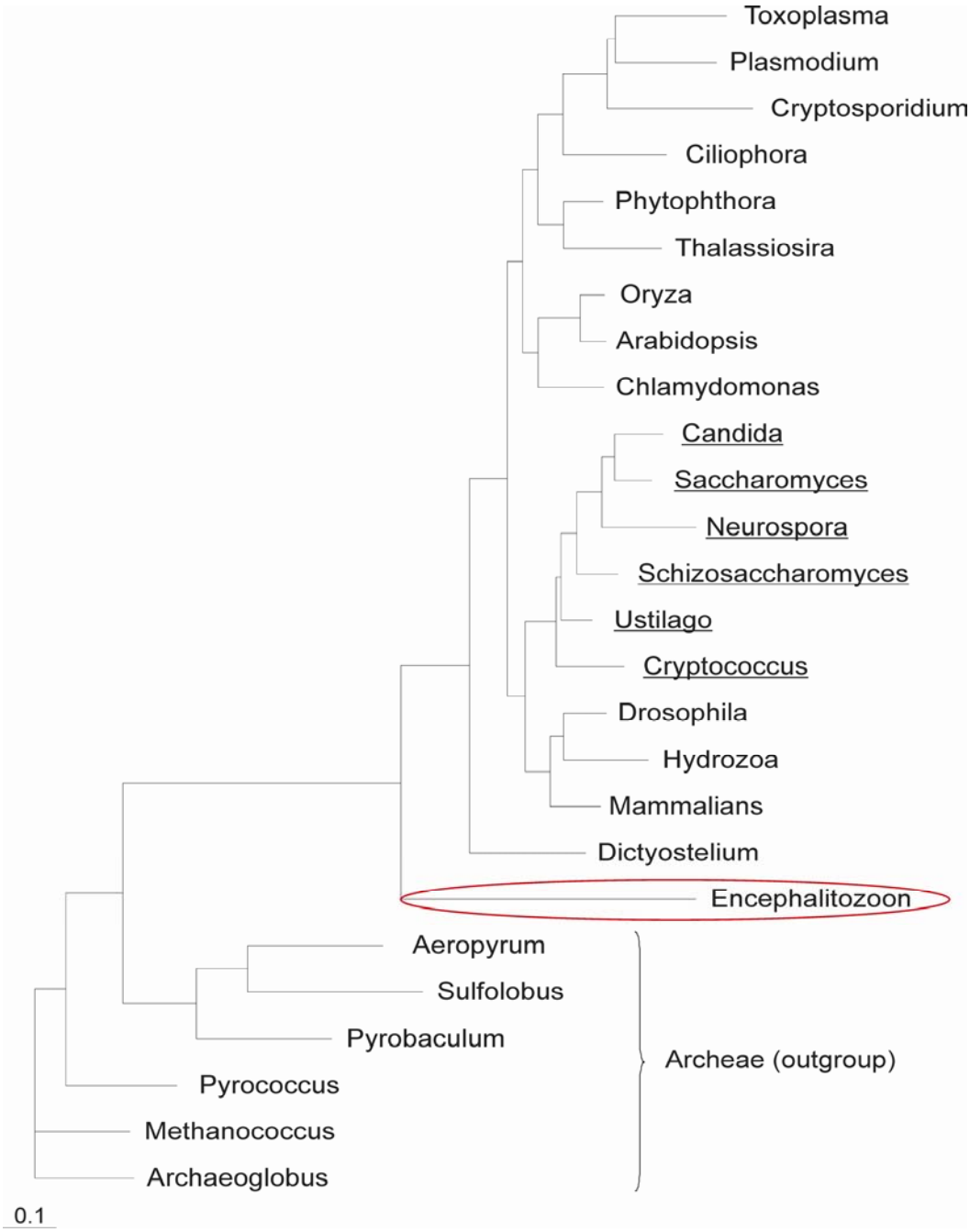


Figure 4.4 – An example of a gene (A-cct) where Microsporidia has the longest terminal branch on the tree (and is the basal eukaryote). Fungal species are underlined.

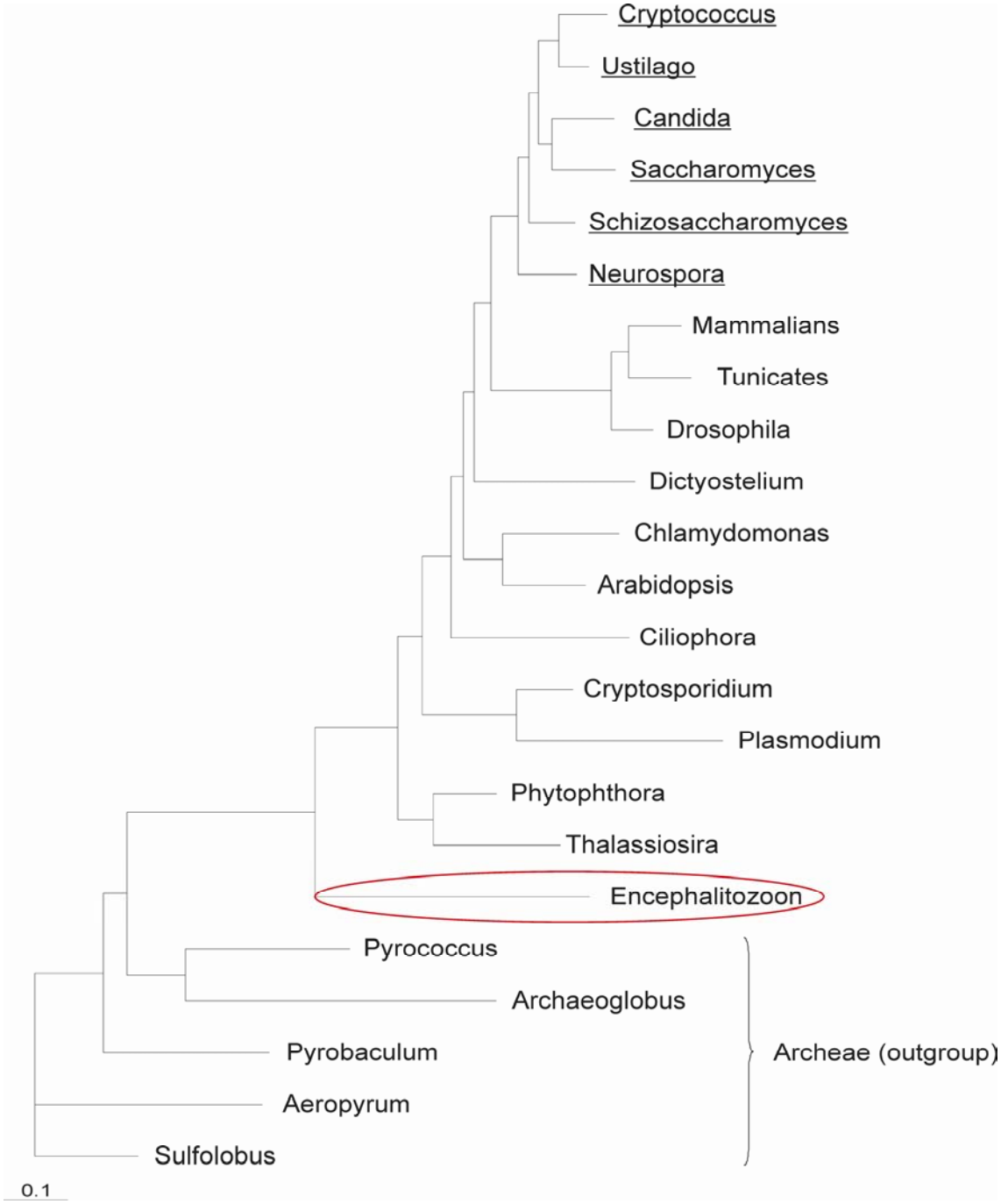


Figure 4.5 – An example of a gene (D-mcm) where Microsporidia is basal on the eukaryote tree (but is not the longest terminal branch). Fungal species are underlined.

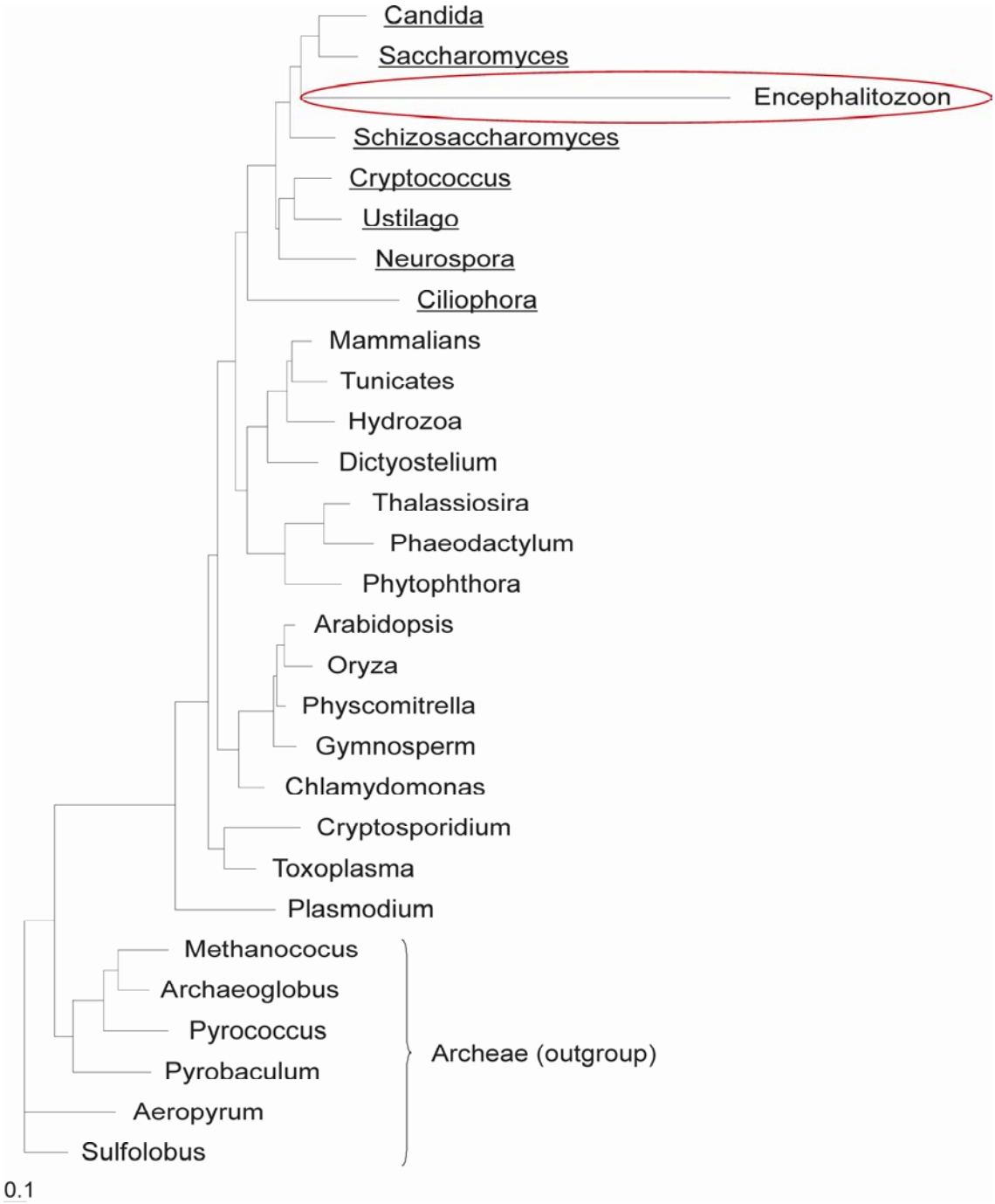


Figure 4.6 – The gene (E-psma) for which Microsporidia groups within Fungi and the *Q* matrix for Microsporidia is significantly different from the others. Fungal species are underlined.

4.4.3 A change in process can cause long branches

Our test can detect changes in the substitution process from which an increase in Prad in long branches can be inferred. Such changes are not accounted for by the models

currently used for tree reconstruction. Phylogenetic methods are known to fail in reconstructing the correct phylogeny when the assumptions they encompass are violated by the actual biochemical evolutionary process underlying gene and protein evolution [28, 29]. Changes in the substitution process are therefore expected to cause model misspecification when a single substitution matrix is applied to the entire dataset.

To demonstrate the effect of this type of model misspecification on tree reconstruction and branch estimation we simulated sequences, with an introduced change in the substitution process (see material and methods), along the tree shown in Figure 4.2. Tree reconstruction was then done using a single substitution matrix. Figure 4.7 shows the unrooted tree with the average estimated branch lengths. While most branches are estimated correctly, the branch length leading to taxon T4 which evolved under a different substitution process (equal relative rates vs. Dayhoff matrix) is mistakenly estimated as one and a half times its actual (simulated) length.

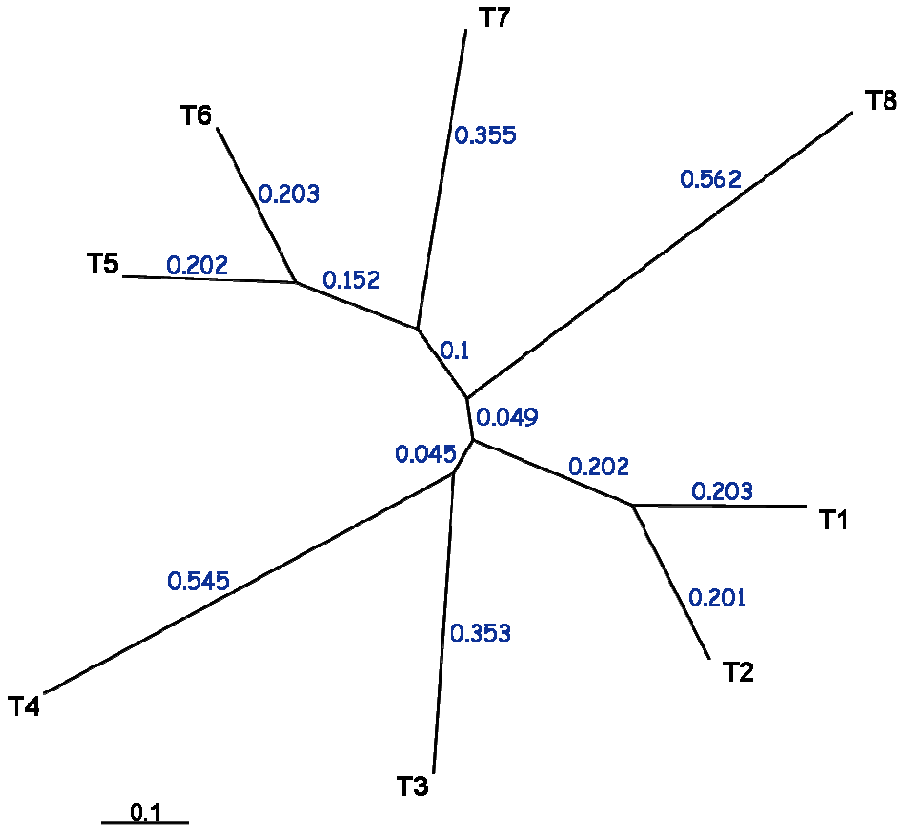


Figure 4.7 – The tree with average branch lengths as estimated for the 100 datasets using the Dayhoff substitution matrix (compare edge lengths to those in Figure 4.2).

4.5 **Conclusions**

We introduced and validated (Figure 4.3) a non-parametric bootstrap test to detect changes in evolutionary processes. When applying the test to the Microsporidia data we found that increased proportions of radical substitutions are linked with long terminal branches for the Microsporidia lineage in about 90% of the cases, and with a basal eukaryote position for the Microsporidia branch in about 84% of the cases. Our results support the conclusions from earlier studies that the basal placement of Microsporidia is the result of a LBA artifact [19, 20, 21, 22, 23, 24, 25], where the long Microsporidia branch is attracted to the long branches leading to the Archaea outgroup. However, we found that the long Microsporidian branch is likely to be an erroneous estimation due to an increased proportion of radical substitutions on that branch, rather than an increased evolutionary rate. Additional work needs to be done to investigate whether sampling error can bias the bootstrap test (through the rate matrix estimation) towards identification of change in process in long branches.

Using simulated data we found that tree reconstruction under a single substitution matrix, when the sequences evolve under multiple matrices (a change in the evolutionary process occurs), can cause erroneous estimate of branch lengths (Figure 4.2 and Figure 4.7). One can rank the sites in the alignment according to their deviation from a homogeneous process (i.e. the average rate matrix) and assign a weight for each site (where the weight can be larger or equal to zero) prior to the phylogenetic reconstruction. This procedure would be similar to the slow-fast (S-F) method [30] currently used for "fast evolving sites"; however rather than removing (or down-weighting) sites according to their overall rate of substitution, sites are excluded (or down-weighted) if they exhibit change in the substitution process. Although we do not advocate the use of such a method as a means of improving phylogenetic accuracy, it might be useful as an exploratory method to determine the effect of the non-homogeneous substitution process on the phylogenetic analysis. One way of dealing with such changes in the evolutionary process may be the use of a general Markov model [31, 32], with several rate matrices, for phylogenetic inference where different

matrices are applied to edges on the path to aberrant taxa. Change in the substitution process, as reported here, is expected to be correlated with change in the proportion of variable sites (as both are manifests of changing evolutionary constraints) which can cause model misspecification and LBA [33, 34 (Chapter 3 in this thesis)].

4.6 **Acknowledgements**

We thank Tal Dagan, Giddy Landan, Masami Hasegawa, and Tal Pupko for valuable discussions. LSG thanks Klaus Schliep for his help with R, and Barbara Holland, David Penny, and Mike Hendy for their comments on this manuscript.

4.7 References

1. Felsenstein J. 1978. Cases in Which Parsimony or Compatibility Methods Will Be Positively Misleading. Syst Zool 27:401-410.
2. Hendy MD and Penny D. 1989. A Framework for the Quantitative Study of Evolutionary Trees. Syst Zool 38:297-309.
3. Bergsten J. 2005. A review of long-branch attraction. Cladistics 21:163-193.
4. Soltis DE, Albert VA, Savolainen V, Hilu K, Qiu YL, Chase MW, Farris JS, Stefanovic S, Rice DW, Palmer JD and Soltis PS. 2004. Genome-scale data, angiosperm relationships, and 'ending incongruence': a cautionary tale in phylogenetics. Trends Plant Sci 9:477-483.
5. Harrison GL, McLenachan PA, Phillips MJ, Slack KE, Cooper A and Penny D. 2004. Four new avian mitochondrial genomes help get to basic evolutionary questions in the late Cretaceous. Mol Biol Evol 21:974-983.
6. Lockhart PJ and Cameron SA. 2001. Trees for bees. Trends Ecol Evol 16:84-88.
7. Clements KD, Gray RD and Howard Choat J. 2003. Rapid evolutionary divergences in reef fishes of the family Acanthuridae (Perciformes: Teleostei). Mol Phylogenet Evol 26:190-201.
8. Lin YH, McLenachan PA, Gore AR, Phillips MJ, Ota R, Hendy MD and Penny D. 2002. Four new mitochondrial genomes and the increased stability of evolutionary trees of mammals from improved taxon sampling. Mol Biol Evol 19:2060-2070.
9. Philippe H. 1997. Rodent Monophyly: Pitfalls of Molecular Phylogenies. J Mol Evol 45:712-715.
10. Philippe H and Laurent J. 1998. How good are deep phylogenetic trees? Curr Opin Genet Dev 8:616-623.
11. Holland BR, Penny D and Hendy MD. 2003. Outgroup misplacement and phylogenetic inaccuracy under a molecular clock - A simulation study. Syst Biol 52:229-238.

12. Shavit L, Penny D, Hendy MD and Holland BR. 2007. The Problem of Rooting Rapid Radiations. Mol Biol Evol 24:2400-2411.

13. Liu Y, Leigh JW, Brinkmann H, Cushion MT, Rodriguez-Ezpeleta N, Philippe H and Lang BF. 2009. Phylogenomic Analyses Support the Monophyly of Taphrinomycotina, including Schizosaccharomyces Fission Yeasts. Mol Biol Evol 26:27-34.

14. Brinkmann H, van der Giezen M, Zhou Y, Poncelin de Raucourt G and Philippe H. 2005. An empirical assessment of long-branch attraction artefacts in deep eukaryotic phylogenomics. Syst Biol 54:743-757.

15. Dickerson RE. 1971. The structures of cytochrome c and the rates of molecular evolution. Molecular Evolution 1:26-45.

16. Goldberg AV, Molik S, Tsaousis AD, Neumann K, Kuhnke G, Delbac F, Vivares CP, Hirt RP, Lill R and Embley TM. 2008. Localization and functionality of microsporidian iron-sulphur cluster assembly proteins. Nature 452:624-628.

17. Leipe DD, Gunderson JH, Nerad TA and Sogin ML. 1993. Small subunit ribosomal RNA+ of Hexamita inflata and the quest for the first branch in the eukaryotic tree. Mol Biochem Parasitol 59:41-48.

18. Kamaishi T, Hashimoto T, Nakamura Y, Nakamura F, Murata S, Okada N, Okamoto K, Shimizu M and Hasegawa M. 1996. Protein phylogeny of translation elongation factor EF-1 alpha suggests microsporidians are extremely ancient eukaryotes. J Mol Evol 42:257-263.

19. Keeling PJ and Doolittle WF. 1996. Alpha-tubulin from early-diverging eukaryotic lineages and the evolution of the tubulin family. Mol Biol Evol 13:1297-1305.

20. Edlind TD, Li J, Visvesvara GS, Vodkin MH, McLaughlin GL and Katiyar SK. 1996. Phylogenetic analysis of beta-tubulin sequences from amitochondrial protozoa. Mol Phylogenet Evol 5:359-367.

21. Keeling PJ and McFadden GI. 1998. Origins of microsporidia. Trends Microbiol 6:19-23.

22. Hirt RP, Healy B, Vossbrinck CR, Canning EU and Embley TM. 1997. A mitochondrial Hsp70 orthologue in Vairimorpha necatrix: molecular evidence that microsporidia once contained mitochondria. Curr Biol 7:995-998.

23. Baldauf SL, Roger AJ, Wenk-Siefert I and Doolittle WF. 2000. A kingdom-level phylogeny of eukaryotes based on combined protein data. Science 290:972-977.

24. Hirt RP, Logsdon JM, Jr., Healy B, Dorey MW, Doolittle WF and Embley TM. 1999. Microsporidia are related to Fungi: evidence from the largest subunit of RNA polymerase II and other proteins. Proc Natl Acad Sci U S A 96:580-585.

25. Inagaki Y, Susko E, Fast NM and Roger AJ. 2004. Covarion shifts cause a long-branch attraction artifact that unites microsporidia and archaebacteria in EF-1 alpha phylogenies. Mol Biol Evol 21:1340-1349.

26. Wang HC, Li K, Susko E and Roger AJ. 2008. A class frequency mixture model that adjusts for site-specific amino acid frequencies and improves inference of protein phylogeny. BMC Evol Biol 8:331.

27. Schliep K. 2009. phangorn: Phylogenetic analysis in R. R package version 0.1-0. http://www.cran.r-project.org/web/packages/phangorn/index.html.

28. Lockhart PJ, Larkum AWD, Steel MA, Waddell PJ and Penny D. 1996. Evolution of chlorophyll and bacteriochlorophyll: The problem of invariant sites in sequence analysis. Proc Natl Acad Sci U S A 93:1930-1934.

29. Felsenstein J. 2004. Inferring phylogenies. Sinauer Associates, Inc., Sunderland, Massachusetts.

30. Brinkmann H and Philippe H. 1999. Archaea sister group of Bacteria? Indications from tree reconstruction artifacts in ancient phylogenies. Mol Biol Evol 16:817-825.

31. Barry D and Hartigan JA. 1987. Statistical Analysis of Hominoid Molecular Evolution. Statistical Science 2:191-207.

32. Jayaswal V, Jermiin, L. S. and Robinson, J. 2005. Estimation of phylogeny using a general Markov model. Evolutionary Bioinformatics 1:62-80.

33. Gruenheit N, Lockhart PJ, Steel M and Martin W. 2008. Difficulties in testing for covarion-like properties of sequences under the confounding influence of changing proportions of variable sites. Mol Biol Evol 25:1512-1520.

34. Shavit Grievink L. 2009. Phylogenetic Tree Reconstruction Accuracy and Model Fit When Proportions of Variable Sites Change Across the Tree. Syst Biol, accepted for publication pending revisions.

4.8 Supplementary Material

4.8.1 Supplementary I

We estimate the normalized instantaneous rate matrix Q as follows.

For two sequences, sequence *x* and sequence *y*, let Π be a diagonal matrix where Π_{ii} is the proportion of sites that have character *i* in sequence *x*. Let F_{ij} denote the proportion of sites that have an *i* in sequence *x* and a *j* in sequence *y*. When calculating these two matrices, we ignore the sites where one or the other has a missing state or a gap. Under the standard model,

(2) $F = \Pi e^{Qt}$

where *Q* is the instantaneous rate matrix and *t* is the divergence time between the sequences. We will assume that *Q* has rate 1, which means that $-trace(\Pi Q) = 1$.

We want to estimate *Q* from *F*. Inverting (1) gives

(3) $Qt = \log(\Pi^{-1} F)$

where log is the matrix log. The log is often difficult to calculate. However, following Tajima[43], we can consider a series expansion to estimate the log. Define *G* = Π - *F* so that

(4) $Qt = \log(\Pi^{-1} F)$

(5) $= \log(I - \Pi^{-1} G)$

$$= -\sum_{k=1}^{\infty} \frac{\left(\Pi^{-1}G\right)^k}{k} \quad (6)$$

$$\approx -\sum_{k=1}^{K} \frac{\left(\Pi^{-1}G\right)^k}{k} \quad (7)$$

for some *K*. As $t \rightarrow 0$, G $\rightarrow$ 0, so the series will converge faster the smaller *t* is.

Q is then estimated by

$$\hat{Q} = \frac{-\sum_{k=1}^{K} \frac{\left(\Pi^{-1}G\right)^k}{k}}{trace\left(\sum_{k=1}^{K} \frac{\Pi\left(\Pi^{-1}G\right)^k}{k}\right)} \quad (8)$$

K=30 was used for normalized instantaneous rate matrices estimation in this chapter.

4.8.2 Supplementary II

Gene	**Longest** +=longest -=not longest	**Position** +=basal -=not basal *=within Fungi	**Change in substitution process** +=change identified -=no change identified	**Increased proportion of radical substitutions** +=increase identified -=no increase identified
A-cct	+	+	+	+
A-psma	+	+	+	+
B-cct	+	+	+	+
B-psma	+	+	+	+
C-l12e	+	+	+	+
C-psma	+	+	+	+
D-psma	+	+	+	+
E-cct	+	+	+	+
EF1-ef1	+	+	+	+
EF2-ef2	+	+	+	+
G-cct	+	+	+	+
G-psma	+	+	+	+
H-psma	+	+	+	+
N-cct	+	+	+	+
T-cct	+	+	+	+
h4	+	+	+	+
if1a	+	+	+	+
if2b	+	+	+	-
l10a	+	+	+	+
l13a	+	+	+	+
l15e	+	+	+	+
orf2	+	+	+	+
rpl14	+	+	+	+
rpl18	+	+	+	+
rpl2	+	+	+	+
rpl26	+	+	+	+
rpl30	+	+	+	+
rpl39	+	+	+	+
rps13	+	+	+	+
rps14	+	+	+	+
rps15	+	+	+	+
rps16	+	+	+	+
rps2	+	+	+	+
rps20	+	+	+	+
rps23	+	+	+	+
rps3	+	+	+	+
rps3a	+	+	+	+
rps4	+	+	+	+
rps5	+	+	+	+
s15a	+	+	+	+
s15p	+	+	+	+
sap40	+	+	+	+
sra	+	+	+	+

Gene	**Longest** +=longest -=not longest	**Position** +=basal -=not basal *=within Fungi	**Change in substitution process** +=change identified -=no change identified	**Increased proportion of radical substitutions** +=increase identified -=no increase identified
wrs	+	+	+	+
A-nsf2	+	+	-	-
J-nsf1	+	+	-	-
J-psma	+	+	-	-
rpl5	+	+	-	-
E-psma	+	*	+	+
C-rpo	+	*	-	-
A-rla2	+	-	+	+
B-pace2	+	-	+	+
B-rrp46	+	-	+	+
D-cct	+	-	+	+
F-psma	+	-	+	+
N-psmb	+	-	+	+
RF3-ef1	+	-	+	+
Z-cct	+	-	+	+
mra1	+	-	+	+
pace5	+	-	+	+
rpl17	+	-	+	+
rpl9	+	-	+	+
rps19	+	-	+	+
s27e	+	-	+	+
D-l12e	+	-	-	-
l28e	+	-	-	-
A-rpl7	-	+	+	+
D-mcm	-	+	+	+
I-psma	-	+	+	+
if2g	-	+	+	+
l11b	-	+	+	+
l19e	-	+	+	+
l37a	-	+	+	-
rpl1	-	+	+	+
rpl11	-	+	+	+
rpl27	-	+	+	+
rpl3	-	+	+	-
rpl34	-	+	+	-
rpl44	-	+	+	+
rps11	-	+	+	+
rps17	-	+	+	+
rps8	-	+	+	+
B-rpo	-	+	-	-
C-mcm	-	+	-	-
I-nsf1	-	+	-	-
K-nsf1	-	+	-	-
L-nsf1	-	+	-	-
rpl32	-	+	-	-
srp54	-	+	-	-

Gene	**Longest** +=longest -=not longest	**Position** +=basal -=not basal *=within Fungi	**Change in substitution process** +=change identified -=no change identified	**Increased proportion of radical substitutions** +=increase identified -=no increase identified
vata	-	+	-	-
B-mcm	-	*	-	-
fibri	-	*	-	-
vatb	-	*	-	-
xpb	-	*	-	-
A-rpo	-	-	+	+
B-l12e	-	-	+	-
C-pace2	-	-	+	+
E-mcm	-	-	+	+
if6	-	-	+	+
l10b	-	-	+	+
l35a	-	-	+	+
l37e	-	-	+	+
rpl10	-	-	+	+
rpl25	-	-	+	+
A-mcm	-	-	-	-
A-rad51	-	-	-	-
B-rpl24	-	-	-	-
F-mcm	-	-	-	-
F-nsf2	-	-	-	-
G-nsf1	-	-	-	-
K-psmb	-	-	-	-
L-psmb	-	-	-	-
M-nsf1	-	-	-	-
M-psmb	-	-	-	-
crfg	-	-	-	-
if2p	-	-	-	-
rf1	-	-	-	-
rpl21	-	-	-	-
rps29	-	-	-	-
rps6	-	-	-	-
srs	-	-	-	-
tfiid	-	-	-	-

Chapter 5

The Enigma of Mesostigma

Manuscript in preparation; in collaboration with Barbara Holland, David Penny, Mike Hendy, and Peter Lockhart.

5.1 **Abstract**

Mesostigma is a fresh-water green alga. It is an isolated taxon for which contradictory phylogenetic relationships have been inferred; while some analyses placed it as sister to all green plants, others have placed it within the Streptophyta lineage. Previous work suggested that the basal placement of Mesostigma is a result of a long-branch attraction artifact due to poor taxon sampling. Using mitochondrial amino acid sequence data, we show that in this case site sampling (and in particular the treatment of missing data) is just as important a factor for tree reconstruction accuracy. Nevertheless, when a 13-taxon sample was used the results were less sensitive to model choice in comparison to an 8-taxon sample. We found that recreating the long-branch attraction observed for the 8-taxon sample in simulated data is difficult. This is likely to be a result of biochemical properties of proteins that are unaccounted for in current models, such as the low number of amino acid character states per site which we observed in the real dataset.

5.2 **Introduction**

Green plants comprise two major phyla: Streptophyta (land plants and their closest green algal relatives) and Chlorophyta (other extant green, mostly aquatic, algae) [1]. Molecular studies have revealed phylogenetic relationships among major green plant lineages [1, 2]. Nevertheless, some incongruence still remains. One such case is that of *Mesostigma viride* (common name Mesostigma), the only known member of Mesostigmatales [1]. Because it is such an isolated taxon, with a lineage which is likely to extend back a billion years, it is unsurprising that Mesostigma is difficult to place accurately as it is expected to be prone to long-branch attraction (LBA; where two non-sister long-branch lineages are incorrectly grouped together in a phylogeny) [3, 4]. Mesostigma is a fresh-water, unicellular, green, scaly bi-flagellate. It was first classified as a member of Chlorophyta (belonging to its earliest diverging lineage Prasinophyceae) [5]. More recently, some phylogenetic analyses have placed it as basal to all other greens [6, 7, 8, 9, 10], before the split of Streptophyta and Chlorophyta, while others [2, 5, 11, 12, 13, 14, 15] have suggested that it is the earliest divergence within Streptophyta.

Mesostigma represents the earliest divergence of the Streptopyta in phylogenies based on large multigene analyses of nuclear, plastid, and mitochondrial datasets [11], four genes (nuclear 18S rRNA gene, chloroplast atpB and rbcL, and mitochondrial nad5) [5], three genes (nuclear 18S rDNA, chloroplast atpB and rbcL) [12], and in trees based on 18S rDNA [13], actin genes [14], chloroplast genes [15], and chloroplast genomes [2]. In contrast, phylogenies based on other datasets of either multiple mitochondrial genes [6] or multiple chloroplast genes [7, 8, 9, 10] placed Mesostigma as basal to Streptophyta and Chlorophyta.

Other evidence is also conflicting. All Streptophyta were shown to have elongation factor-1 alpha (EF-1α), whereas Mesostigma possess elongation factor-like (EFL) [16]. Mesostigma's photosynthetic pigment composition also supports its basal position [17]. On the other hand, Mesostigma was shown to share more ESTs (i.e. expressed sequence

tag; see [18] for more information) with land plants than with the chlorophyte *Chlamydomonas reinhardtii* [19]. Nevertheless, the within Streptophyta position now seems to be generally accepted [20]. Whatever its final phylogenetic position, as one of the most primitive green algae, Mesostigma is essential for the understanding of the evolution of green plants.

The specific causes for the incongruence between different phylogenetic studies, regarding the position of Mesostigma, are unclear, and are the topic of the present study. Turmel et al. [6] analyzed a mitochondrial dataset of concatenated protein sequences based on 19 genes (4,139 amino acid positions) from 8 taxa (Mesostigma, 4 green algae, and 3 red algae). The resulting trees (using maximum likelihood and distance trees with the JTT model of amino acid substitution, as well as maximum parsimony) placed Mesostigma as a basal green alga before the divergence of Streptophyta and Chlorophyta, with strong (100%) bootstrap support. Rodríguez-Ezpeleta et al. [11] analyzed an extended mitochondrial dataset of 33 concatenated protein sequences from 13 taxa (8 green plants, and 4 red algae and a jakobid flagellate as outgroups). The dataset used by Turmel et al. [6] is a subset, both in the taxa and sites, of that used by Rodríguez-Ezpeleta et al. [11]. The resulting maximum likelihood tree (using the WAG+Γ+F model) placed Mesostigma as the earliest divergence within Streptophyta. This tree was only weakly supported (and was not found by maximum parsimony). Nonetheless, because this tree was congruent with their analysis of nuclear data and with previous single gene phylogenies, they concluded that the placement of Mesostigma as basal to Streptophyta and Chlorophyta was an artifact [11]. After adjusting their dataset to that used by Turmel et al. [6], the authors suggested that the likely reason for the discrepancy is poor taxon sampling combined with failure to account for rate heterogeneity among sites and that the number of sites used was less important.

In a recent study, we have demonstrated [21] (Chapter 3 in this thesis), using nucleotide sequences, that lineage-specific proportions of variable sites can cause model misspecification which can lead to LBA. Our findings show that the rates-across-sites model (although it is not based on any specific biochemical process) can, at least

partially, account for changes in the proportion of variable sites (Pvar), and that an increased taxon sample can improve tree reconstruction accuracy under these conditions. Lineage-specific proportions of variable sites might therefore explain the observations of Rodríguez-Ezpeleta et al. [11], where increased taxon sampling and/or the use of rates-across-sites model resulted in seemingly more accurate phylogenies. The aim of the study presented here is to gain insight into the discrepancy between different studies regarding the positioning of Mesostigma. We focus on the dataset used by Rodríguez-Ezpeleta et al. [11] and the subset of taxa from this dataset which was used by Turmel et al. [6]. The study presented here was limited by the lack of information regarding the sequence alignment (such as which proteins were used and what sites comprise which protein).

5.3 Materials and Methods

5.3.1 Mesostigma dataset

We used the mitochondrial dataset from Rodríguez-Ezpeleta et al. (supplementary material) [11], containing 33 proteins, a total of 6622 amino acid positions, from 13 taxa (8 green plants: Mesostigma, 5 Streptophyta, 2 Chlorophyta, and 4 red algae and a jakobid as outgroups).

5.3.2 Phylogenetic analysis

The best-fit model was determined using the program ProtTest [22] with the Akaike Information Criterion [23] starting with a BioNJ tree and optimizing topology and branch lengths. The maximum likelihood (ML) tree and branch lengths were also estimated using the program PhyML v. 3.0 [24] with the JTT, WAG, and CpREV models and all combinations of +I (constant proportion of invariable sites), +G (gamma distribution for rates across sites), and +F (empirical character frequencies).

5.3.3 Simulations

Sequence data was generated using LineageSpecificSeqgen [25] (Chapter 2 in this thesis). One hundred datasets of 6622 amino acid positions were simulated using either the CpREV or WAG model with the corresponding rooted trees (using Reclinomonas as an outgroup) and branch lengths (identical topology, branch lengths differ slightly i.e. <0.0072). The proportion of invariable sites (I) at the root of the tree was varied from 0 to 0.4 in steps of 0.1. Change in the proportion of variable sites was introduced on the Meosotigma lineage (see Figure 5.1) where a fraction, $Pvar^{+}$ = (0, 0.1, 0.2, 0.3, 0.4, 0.5, 0.6, 0.7, 0.8, 0.9, 1), of the invariable sites were set to be variable.

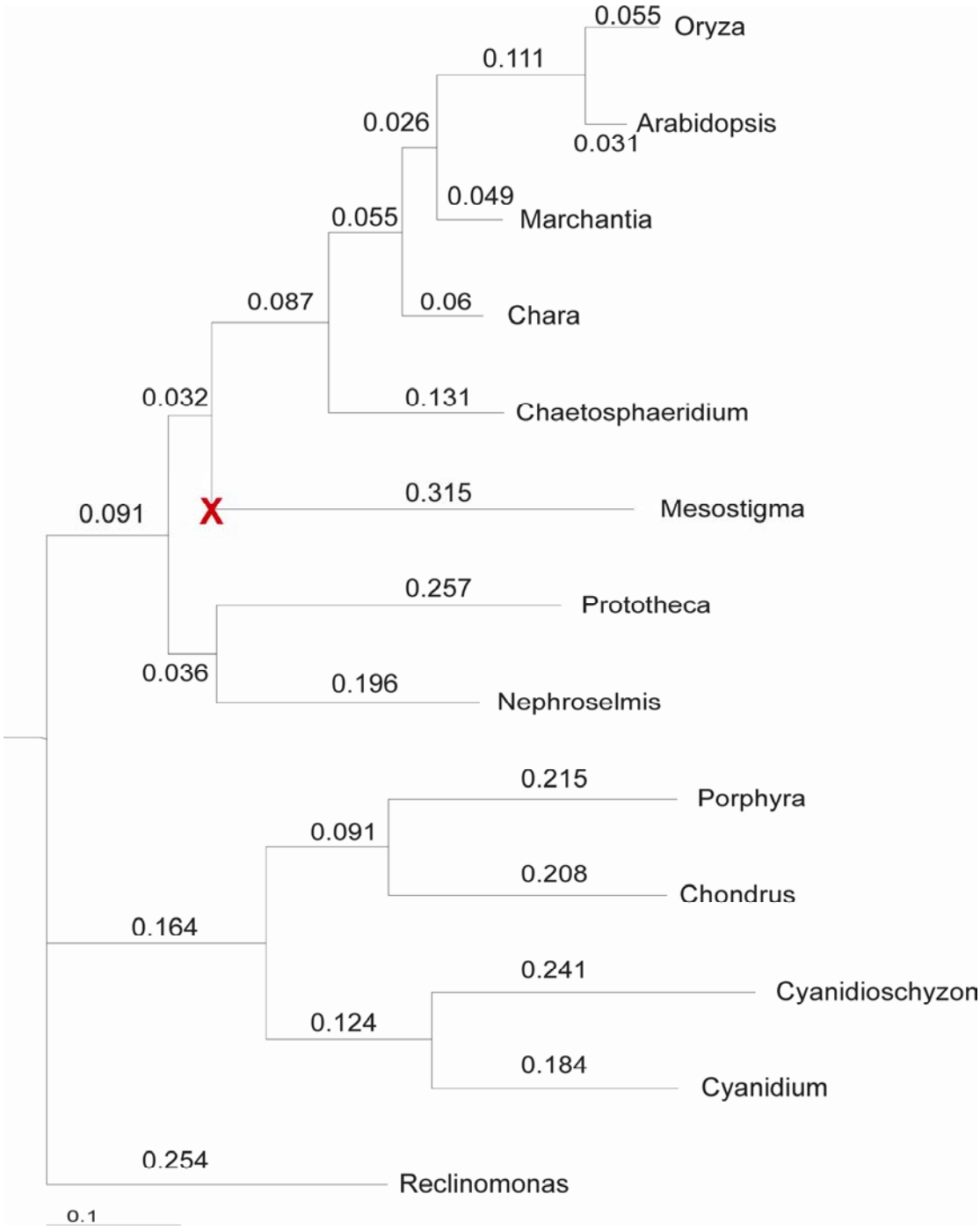


Figure 5.1 - The tree used for simulations with branch lengths according to the WAG model (CpREV branch lengths differ slightly i.e. <0.0072). The proportion of invariable sites at the root was varied from 0 to 0.4, in steps of 0.1. Change in the proportion of variable sites was introduced on the Mesostigma lineage (marked with X), where a fraction, Pvar+ = (0, 0.1, 0.2, 0.3, 0.4, 0.5, 0.6, 0.7, 0.8, 0.9, 1), of the invariable sites were set to be variable.

5.4 Results and Discussion

5.4.1 Tree estimation differences between the 13-taxa dataset and the 8-taxa subset

The models that were used in the two original studies [6, 11] differ from each other (WAG and JTT, respectively). Therefore, tree estimation with these models cannot be compared directly for the 13-taxa and 8-taxa datasets. For that reason, we estimated the phylogeny using each of these models, as well as the best-fit model selected by the program ProtTest [22], with all combinations of +I, +G, and +F. Unexpectedly, the CpREV+Γ+I+F model was selected as the best-fit model. This model selection is very surprising as the sequences in the dataset are derived from the mitochondria, whereas the CpREV model is based on chloroplast datasets. Other models (for example MtREV) which are based on mitochondrial datasets and therefore intuitively should be a better fit for the dataset were included in the set of models that were tested, but were not selected. The model used in Rodríguez-Ezpeleta et al. [11] (WAG+G+F) was the fourth-best model (with ΔAIC, the difference in AIC score from the best-fit model, of 886.01). The model used by Turmel et al. [6] (JTT) had a ΔAIC of 11271.59 and was one of the worst-fit models, even when the dataset was reduced to their taxon sampling (ΔAIC of 7394.48). However it is important to note that the sites in the dataset of Turmel et al. are a subset of the sites in our dataset.

In a recent study [21] we found that the best-fit model selected using a relative test might be inaccurate in tree reconstruction, and that an absolute model-fit test (such as those described by Goldman [26] and Bollback [27]) may provide a better prediction for phylogenetic reconstruction accuracy. Unfortunately, the absolute model-fit test in a ML framework [26] is not yet implemented in a readily available program. We therefore carried out the ML analysis with each of these three models (JTT, WAG, CpREV) and all possible combinations of +I, +G, and +F.

Because the 8- and 13-taxa datasets also (naturally) differ in the number of sites containing gaps and missing data, and since missing data may affect tree reconstruction accuracy [28, 29] we considered 5 different combinations of taxon and site sampling: (1) the original 13-taxa dataset, (2) The original dataset reduced to 8-taxa, (3) The 13-taxa dataset with gaps and missing sites removed, (4) the reduced 8-taxa dataset with gaps and missing sites removed, and (5) The 13-taxa dataset with missing sites and gaps removed, and then reduced to the 8-taxon sample. The positioning of Mesostigma in the resulting phylogenies is shown in Table 5.1.

Table 5.1 – The positioning of Mesostigma in trees estimated using three different models (JTT, WAG, CpREV) and combination of +I, +G, and +F. 'S' = within Streptophyta, 'B' = basal to green plants. The best-fit model, found using ProtTest, for each of the settings is marked with a *.

Model	**Original 13-taxa dataset** (6622 positions)	**13-taxa dataset gaps and missing sites removed** (1948 positions)	**8-taxa dataset** (6622 positions)	**8-taxa dataset gaps and missing sites removed** (3910 positions)	**8-taxa dataset gaps reduced from the 13-taxa set with gaps and missing sites removed** (1948 positions)
JTT	S	S	B	B	S
JTT+F	S	S	S	B	S
JTT+I	S	S	B	B	S
JTT+I+F	B	S	B	B	S
JTT+G	S	S	B	S	S
JTT+G+F	S	S	S	S	S
JTT+I+G	S	S	S	S	S
JTT+I+G+F	S	S	S	S	S
WAG	S	S	B	B	S
WAG+F	S	S	B	B	S
WAG+I	S	S	B	B	S
WAG+I+F	S	S	B	B	S
WAG+G	S	S	B	S	S
WAG+G+F	S	S	S	S	S
WAG+I+G	S	S	S	S	S
WAG+I+G+F	S	S	S	S	S
cpREV	S	S	B	B	S
cpREV+F	S	S	B	B	S
cpREV+I	S	S	B	B	S
cpREV+I+F	S	S	B	B	S
cpREV+G	S	S	B	S	S
cpREV+G+F	S	S	S*	S	S
cpREV+I+G	S	S	S	S	S
cpREV+I+G+F	S*	S*	S	S*	S*

The results in Table 5.1 clearly show that, in this case, site sampling is as important as taxon sampling. Thus both site sampling and taxon sampling are important. Nevertheless, the 13-taxa dataset is much more robust with respect to the model used. Rodríguez-Ezpeleta et al. [11] found that the number of positions used was less important as long as rate heterogeneity among sites was modeled. In contrast, our results show that when rate heterogeneity is the only estimated distribution (I and F are not estimated), the tree estimation for the 8-taxa dataset is sensitive to site sampling. Exclusion of sites with missing data results in the positioning of Mesostigma within Streptophyta (as is the case for the 13-taxa dataset) while inclusion of sites with missing data results in its basal positioning.

We also found that removing sites with missing data from the complete (13-taxa) alignment and then reducing the dataset to the subset of 8-taxa results in the placement of Mesostigma within Streptophyta, regardless of model choice (column 5 in Table 5.1). The number of incomplete characters in these datasets is much larger than the number of complete characters, the effect of this is unknown [28]. Programs used to estimate the likelihood of phylogenetic trees often treat missing characters by summing over the probabilities of all possible characters. Although handling missing characters in this way may be intuitive from a statistical point of view, our knowledge about sequence evolution suggests that a more reasonable method would be to calculate the probability of a character using the known characters at the same site (assigning higher probabilities for characters that already exist at the site). Nevertheless, estimating the probability of a given amino acid at a site based on the small sample of observed characters at that site might have undesirable statistical properties (and in some cases may lead to inconsistency).

The strong effect site sampling has on tree reconstruction for the 8-taxon dataset, led us to examine the likelihood (using the WAG model) of each of the two competing trees (shown in Figure 5.2) for each site. We considered the 13-taxon dataset with and without removal of gaps and missing sites and the 8-taxon dataset reduced from each of these. The results are summarized in Table 5.2.

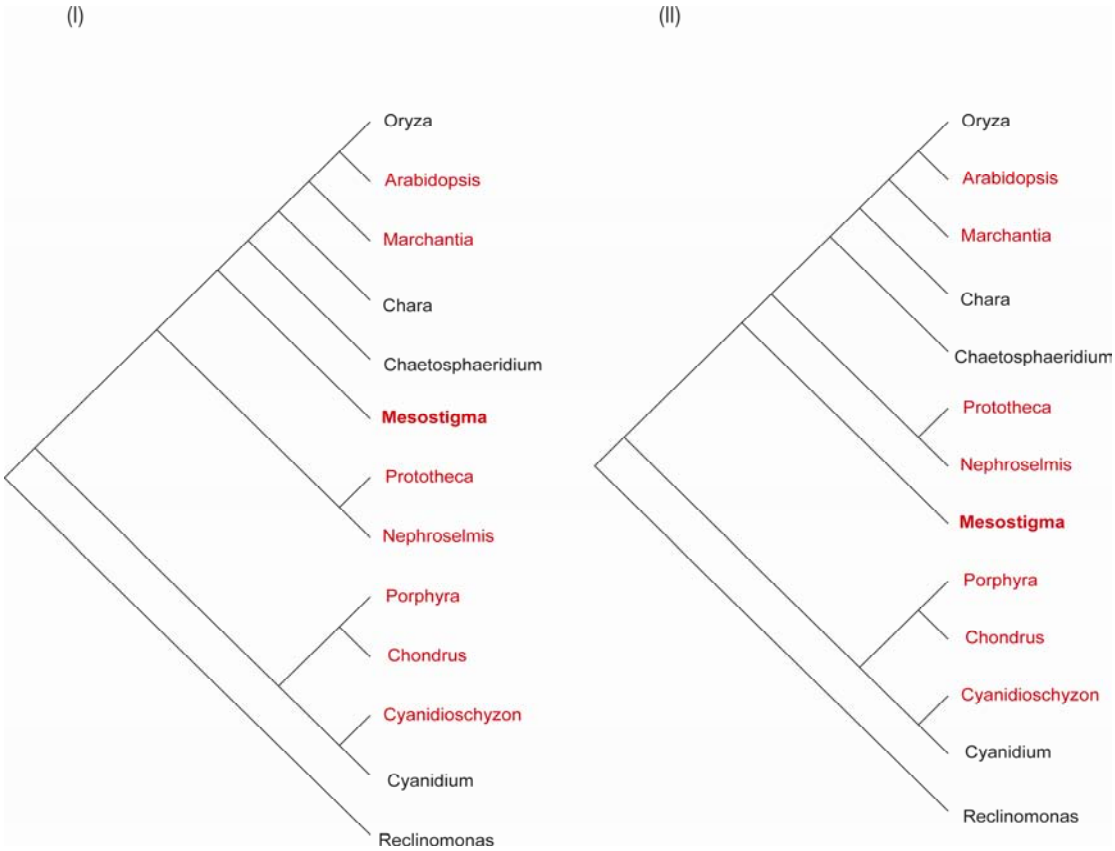


Figure 5.2 – The two competing maximum likelihood trees. In tree (I) Mesostigma is positioned within the Streptophyta (S), whereas in tree (II) Mesostigma is basal to both Streprophyta and Chlorophyta (B). The 8 taxa which are included in the 8-taxon dataset are marked in red.

Table 5.2 – Summary of site likelihoods (using the WAG model) for the 8- and 13- taxon datasets, with and without the removal of gaps and missing data. 'S' = within Streptophyta, 'B' = basal to green plants.

Dataset	Gaps and missing data	Mesostigma position in the ML tree	# sites supporting the within Streptophyta position	# sites supporting the basal position	Total number of sites	Averaged difference in likelihood per site between the ML tree and the alternative tree (rounded to 4th decimal place)
13-taxon	included	S	2506	**4116**	6622	0.0017
	excluded	S	**1457**	491	1948	0.0134
8-taxon	included	B	2768	**3854**	6622	0.0074
	excluded	S	**1510**	438	1948	0.0138

Interestingly, we found that for the 13-taxon dataset when gaps and missing data are included in the analysis ~62% of the sites supported the basal positioning of Mesostigma (B), while Mesostigma is placed within Streptophyta (S) in the ML tree. The percentage of sites supporting the within-Streptophyta positioning increases when gaps and missing data are removed.

For the 8-taxon dataset with gaps and missing data included ~58% of the sites support the basal positioning of Mesostigma (the ML tree for this dataset). However, with gaps and missing sites excluded, ~78% of the remaining sites support the placement of Mesostigma within Streptophyta, the ML tree in this case. These results, together with the low bootstrap support found by Rodríguez-Ezpeleta et al. [11], suggest that site sampling is an extremely important and problematic factor in this case. A larger sequence length is required to infer the position of Mesostigma with confidence.

5.4.2 Change in Pvar as a possible cause for the discrepancy

Addressing the question of whether a change in Pvar in the Mesostigma lineage is the cause for the incongruent results regarding its position is not straightforward. Pvar estimation is difficult and might be unreliable. This is because sites can be variable (free to accept substitutions) but invariant (no substitutions are observed in the sampled group of taxa). The estimation of Pvar is directly affected by the number of taxa sampled; the accuracy is expected to increase with the size of the taxon sample. From a biochemical perspective, Pvar is expected to change over time. This is due to variations in the structural and functional constraints that are acting on the sequences [30, 31]. Lineage-specific proportions and positions of variable sites make this problem even more difficult [32]. We therefore chose to use simulations, as an initial step, to test whether the basal positioning of Mesostigma can be constructed when a change in Pvar is introduced in the Mesostigma lineage.

We conducted simulations mimicking the case of Mesostigma, generating 13-taxon datasets using the empirical character frequencies, and each of the three models (JTT, WAG, and CpREV) and their respective ML tree. A change in Pvar was introduced in

the Mesostigma lineage as shown in Figure 5.1. The datasets were then reduced to the 8-taxon sample and tree reconstruction accuracy using each of the models was evaluated. For the tree and parameter combinations used, the introduced change in Pvar did not affect the tree estimation accuracy and the within Streptophyta position for Mesostigma was reconstructed 100% of the time. However, Brikmann et al. [33] found that phylogenetic methods tend to be more robust in relation to LBA when estimating phylogenies from simulated datasets in comparison to real datasets. Wang et al. [34] have shown that the number of different amino acid character states per site in real datasets is much lower than that in simulated dataset. For simulated amino acid datasets, using any of the empirical models, all amino acid substitutions are possible with some probability. In real data however, the process is much more complex than that captured by the standard models; the constraints acting on the sequence determine the possible types of substitutions that may occur at any given site and the probability of many substitutions may be zero, in which case only a very small number of different characters will be observed at the site.

To test whether this difference between the real dataset and our simulated datasets can be observed we calculated the frequencies of different numbers of unique amino acid character states per site for both the real (Figure 5.3) and the simulated data (Figure 5.4 and Figure 5.5) using our own python code (for simulated data the numbers of unique characters per site were averaged over the 100 datasets). The results show that the patterns for the real data are different from those in our simulated data. The higher the Pinv and the higher the number of invariable sites that become variable in the Mesostigma lineage, the closer the patterns of the simulated data are to those of the real data.

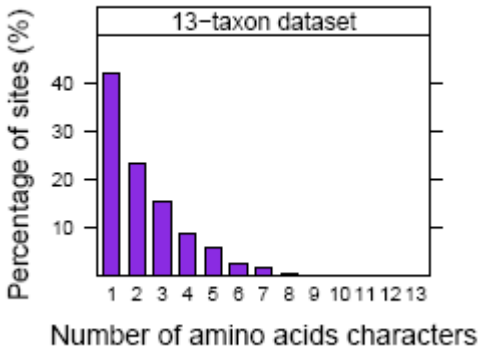


Figure 5.3 – The average number of amino acid character states per site for the 13-taxon dataset.

Figure 5.4 – The average number of amino acid character states per site in simulated 13-taxon datasets under the WAG model. I is the proportion of invariable sites, while Pvar refers to the proportion of invariable sites that are set to be variable. Similar results were obtained for simulations using the CpRev model (results not shown). The results in this figure should be compared to Figure 5.3. See Figure 5.5 for additional parameters (Pvar = 0.6-1).

Figure 5.5 - The average number of amino acid character states per site in simulated 13-taxon datasets under the WAG model and for Pvar=0.6 to 1.0. Similar results were obtained for simulations using the CpRev model (results not shown). The results in this figure should be compared to Figure 5.3. See Figure 5.4 for additional parameters (Pvar = 0-0.5).

Even when the best-fit model is used (CpREV+I+G+F), with the respective estimated parameters, the average number of amino acids character states per site (Figure 5.6) is higher than that in the real data (Figure 5.3). In particular, the number of invariant

(constant) sites in the real data is much higher than that of the simulated data, suggesting that Pinv is underestimated in the ML analysis of the real data.

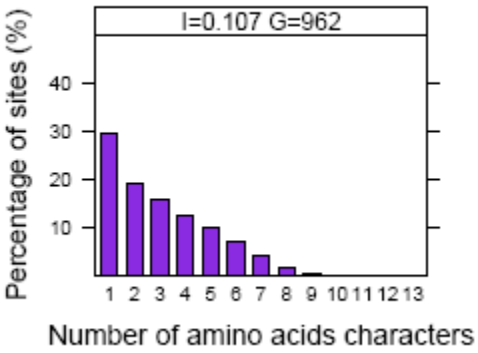


Figure 5.6 - The average number of amino acid characters per site in 13-taxon datasets simulated using the best-fit model (CpREV+I+G+F, with the ML estimates for the tree and parameters).

We also observed differences in the types of substitutions that occurred in the real data versus those of the simulated data (results not shown). While in ~65% of the sites of the real data the amino acids at a site all belong to the same Dayhoff class, for the data simulated under the best-fit model (CpREV+I+G+F) in only ~43% of the sites (on average) do the amino acids at a site all belong to the same Dayhoff class. Even when ignoring the constant sites (~42% in the real data and ~ 30%, on average, in the simulated data) it is clear that the substitutions in the real data are more biochemically conservative than those in the simulated data.

The models used for generating (simulating) datasets are those used for tree estimation. This means that the models used to reconstruct phylogenies are misspecified in that they do not account for the low number of possible types of substitutions in real data. This model misspecification by itself may cause the basal position for Mesostigma. Wang et al. [34] demonstrated that accounting for site-specific preferences of amino acids, using a class frequency mixture model, results in a more accurate phylogenetic estimation for the case of Microsporidia (described in chapter 4), overcoming the LBA artifact. Simulating sequences on the Mesostigma ML tree, with the site-specific amino acid frequencies estimated from the data may result in closer mimicking of the number of

characters per site observed in the real data. Phylogenetic estimation using the usual models can then be applied to test whether this misspecification alone, or combined with change in Pvar, can explain the conflicting results between the 8- and 13-taxa. A simulator (CovTree) which is designed for this purpose exists; unfortunately, it was removed from its author's website (http://morticia.cs.dal.ca/lab_public/?Download:covTREE) and we could not gain access to it even through direct correspondence with the authors.

5.5 **Conclusion**

We found an interesting enigma in that, for 13-taxon dataset used by Rodríguez-Ezpeleta et al. [11], there is a difference between the tree supported by the majority of sites and the tree selected by ML. The majority of sites in this dataset support the basal position for Mesostigma as sister to all green plants. Nevertheless, the sum of log-likelihood values for the within-Streptophyta position was larger, making the within-Streptophyta position more likely and it is therefore selected in ML analyses. These results suggest that the site sample in this dataset is not sufficient to conclude the position of Mesostigma with confidence from this dataset alone.

Our results support the observations of Wang et al. [34] who found that the number of characters at a site in simulated data is significantly larger than that in real data. This is likely to be a result of the underlying biochemical constraints that are acting on the real sequences, limiting the types of possible substitutions in coding sequences, but are unaccounted for by phylogenetic models. An interesting extension for this study would be to compare non-coding regions with current models of substitutions. The CovTree simulator (if it becomes available again) would be useful in testing whether the model misspecification, introduced by not accounting for the small number of characters per site observed in real data, can mislead tree reconstruction.

5.6 **References**

1. Lewis LA and McCourt RM. 2004. Green algae and the origin of land plants. Am J Bot 91:1535-1556.
2. Lemieux C, Otis C and Turmel M. 2007. A clade uniting the green algae Mesostigma viride and Chlorokybus atmophyticus represents the deepest branch of the Streptophyta in chloroplast genome-based phylogenies. BMC Biol 5:2.
3. Felsenstein J. 1978. Cases in Which Parsimony or Compatibility Methods Will Be Positively Misleading. Syst Zool 27:401-410.
4. Hendy MD and Penny D. 1989. A Framework for the Quantitative Study of Evolutionary Trees. Syst Zool 38:297-309.
5. Karol KG, McCourt RM, Cimino MT and Delwiche CF. 2001. The closest living relatives of land plants. Science 294:2351-2353.
6. Turmel M, Otis C and Lemieux C. 2002. The complete mitochondrial DNA sequence of Mesostigma viride identifies this green alga as the earliest green plant divergence and predicts a highly compact mitochondrial genome in the ancestor of all green plants. Mol Biol Evol 19:24-38.
7. Martin W, Deusch O, Stawski N, Grunheit N and Goremykin V. 2005. Chloroplast genome phylogenetics: why we need independent approaches to plant molecular evolution. Trends Plant Sci 10:203-209.
8. Turmel M, Ehara M, Otis C and Lemieux C. 2002. Phylogenetic relationships among streptophytes as inferred from chloroplast small and large subunit rRNA gene sequences. J Phycol 38:364–375.
9. Lemieux C, Otis C and Turmel M. 2000. Ancestral chloroplast genome in Mesostigma viride reveals an early branch of green plant evolution. Nature 403:649-652.
10. Rogers MB, Gilson PR, Su V, McFadden GI and Keeling PJ. 2007. The Complete Chloroplast Genome of the Chlorarachniophyte Bigelowiella natans:

Evidence for Independent Origins of Chlorarachniophyte and Euglenid Secondary Endosymbionts. Mol Biol Evol 24:54-62.

11. Rodriguez-Ezpeleta N, Philippe H, Brinkmann H, Becker B and Melkonian M. 2007. Phylogenetic analyses of nuclear, mitochondrial, and plastid multigene data sets support the placement of Mesostigma in the Streptophyta. Mol Biol Evol 24:723-731.

12. Cocquyt E, Verbruggen H, Leliaert F, Zechman FW, Sabbe K and De Clerck O. 2009. Gain and loss of elongation factor genes in green algae. BMC Evol Biol 9:39.

13. Marin B and Melkonian M. 1999. Mesostigmatophyceae, a new class of streptophyte green algae revealed by SSU rRNA sequence comparisons. Protist 150:399-417.

14. Bhattacharya D, Weber K, An SS and Berning-Koch W. 1998. Actin phylogeny identifies Mesostigma viride as a flagellate ancestor of the land plants. J Mol Evol 47:544-550.

15. Martin W, Rujan T, Richly E, Hansen A, Cornelsen S, Lins T, Leister D, Stoebe B, Hasegawa M and Penny D. 2002. Evolutionary analysis of Arabidopsis, cyanobacterial, and chloroplast genomes reveals plastid phylogeny and thousands of cyanobacterial genes in the nucleus. Proc Natl Acad Sci U S A 99:12246-12251.

16. Noble GP, Rogers MB and Keeling PJ. 2007. Complex distribution of EFL and EF-1alpha proteins in the green algal lineage. BMC Evol Biol 7:82.

17. Yoshii Y, Takaichi S, Maoka T and Inouye I. 2003. Photosynthetic pigment composition in the primitive green alga Mesostigma viride (Prasinophyceae): phylogenetic and evolutionary implications. J Phycol 39:570-576.

18. Adams MD, Kelley JM, Gocayne JD, Dubnick M, Polymeropoulos MH, Xiao H, Merril CR, Wu A, Olde B, Moreno RF and et al. 1991. Complementary DNA sequencing: expressed sequence tags and human genome project. Science 252:1651-1656.

19. Simon A, Glockner G, Felder M, Melkonian M and Becker B. 2006. EST analysis of the scaly green flagellate Mesostigma viride (Streptophyta): implications for the evolution of green plants (Viridiplantae). BMC Plant Biol 6:2.

20. Turmel M, Gagnon MC, O'Kelly CJ, Otis C and Lemieux C. 2009. The chloroplast genomes of the green algae Pyramimonas, Monomastix, and Pycnococcus shed new light on the evolutionary history of prasinophytes and the origin of the secondary chloroplasts of euglenids. Mol Biol Evol 26:631-648.

21. Shavit Grievink L. 2009. Phylogenetic Tree Reconstruction Accuracy and Model Fit When Proportions of Variable Sites Change Across the Tree. Syst Biol, accepted for publication pending revisions.

22. Abascal F, Zardoya R and Posada D. 2005. ProtTest: selection of best-fit models of protein evolution. Bioinformatics 21:2104-2105.

23. Akaike H. 1974. A new look at the statistical model identification. Automatic Control, IEEE Transactions on 19:716-723.

24. Guindon S and Gascuel O. 2003. A simple, fast, and accurate algorithm to estimate large phylogenies by maximum likelihood. Syst Biol 52:696-704.

25. Shavit Grievink L, Penny D, Hendy MD and Holland BR. 2008. LineageSpecificSeqgen: generating sequence data with lineage-specific variation in the proportion of variable sites. BMC Evol Biol 8:317.

26. Goldman N. 1993. Statistical Tests of Models of DNA Substitution. J Mol Evol 36:182-198.

27. Bollback JP. 2002. Bayesian model adequacy and choice in phylogenetics. Mol Biol Evol 19:1171-1180.

28. Wiens JJ. 2006. Missing data and the design of phylogenetic analyses. J Biomed Inform 39:34-42.

29. Huelsenbeck JP. When are Fossils better than Extant Taxa in Phylogenetic Analysis? Syst Zool 40:458-469 CR - Copyright © 1991 Society of Systematic Biologists.

30. Dickerson RE. 1971. The structures of cytochrome c and the rates of molecular evolution. Molecular Evolution 1:26-45.

31. Fitch WM and Markowitz E. 1970. An Improved Method for Determining Codon Variability in a Gene and Its Application to Rate of Fixation of Mutations in Evolution. Biochem Genet 4:579-593.

32. Steel M, Huson D and Lockhart PJ. 2000. Invariable sites models and their use in phylogeny reconstruction. Syst Biol 49:225-232.

33. Brinkmann H, van der Giezen M, Zhou Y, Poncelin de Raucourt G and Philippe H. 2005. An empirical assessment of long-branch attraction artefacts in deep eukaryotic phylogenomics. Syst Biol 54:743-757.

34. Wang HC, Li K, Susko E and Roger AJ. 2008. A class frequency mixture model that adjusts for site-specific amino acid frequencies and improves inference of protein phylogeny. BMC Evol Biol 8:331.

Appendix I

The Problem of Rooting Rapid Radiations

As published in Molecular Biology and Evolution.

Ref: Shavit *L*, Penny D, Hendy MD and Holland BR. 2007. The Problem of Rooting Rapid Radiations. Mol Biol Evol 24:2400-2411.

I.I Abstract

There are many examples of groups (such as birds, bees, mammals, multicellular animals and flowering plants) that have undergone a rapid radiation. In such cases, where there is a combination of short internal and long external branches, correctly estimating and rooting phylogenetic trees is known to be a difficult problem. In this simulation study, we tested the performances of different phylogenetic methods at estimating a tree that models a rapid radiation. We found that maximum-likelihood, corrected- and uncorrected-neighbor-joining and corrected- and uncorrected-parsimony, all suffer from biases towards specific tree topologies. In addition, we found that using a single-taxon outgroup to root a tree frequently disrupts an otherwise correct ingroup phylogeny. Moreover, for uncorrected-parsimony, we found cases where several individual trees (in which the outgroup was placed incorrectly) were selected more frequently than the correct tree. Even for parameter settings where the correct tree was selected most frequently when using extremely long sequences, for sequences of up to 60,000 nucleotides the incorrectly rooted trees were each selected more frequently than the correct tree. For all the cases tested here, tree estimation using a two-taxon outgroup was more accurate than when using a single-taxon outgroup. However, the ingroup was most accurately recovered when no outgroup was used.

I.II Introduction

The problem of tree reconstruction and rooting is known to be challenging, especially in cases of rapid radiations where there is a combination of short and long branches. In particular, long-branch attraction [1, 2, 3] is known to make this problem difficult. Many examples involving birds [4], bees [5], mammals [6], and early divergences of multicellular animals [7, 8], imply that these features are not just of theoretical interest. An example, which has recently highlighted this problem, is the dispute about the rooting of the angiosperms [9, 10, 11, 12]. As pointed out by Lockhart and Penny [13], the basic topology of the angiosperm radiation appears to be star-like (many short internal branches connecting large angiosperm lineages) while the outgroup taxa are relatively distant.

Simulation studies have proven to be useful in evaluating the strengths and weaknesses of phylogenetic methods in tree reconstruction. Previous simulation studies on bifurcating trees show that when internal branches are small relative to external branches even a small misspecification of the substitution model may mislead phylogenetic inference [14, 15]. Holland et al. [16] conducted a simulation study of the performance of the UPGMA, neighbor-joining, maximum parsimony, and maximum likelihood methods, for a five-taxon tree with a symmetric four-taxon ingroup under a molecular clock. That study compared the accuracy of different phylogenetic methods for various sequence lengths, and explored the effectiveness of correcting neighbor-joining for multiple substitutions. Holland et al. [16] also tested the effectiveness of using an outgroup to root a tree and demonstrated some of the problems in reconstructing and rooting trees. They discovered a misleading zone where the tree estimate is consistent (that is, the probability of estimating the correct tree tends to one as the sequence length tends to infinity), but for a wide range of sequence lengths four incorrect trees were each chosen up to twice as frequently as the correct tree. They also established that the inclusion of a distant outgroup, which should join into a short internal branch, frequently disrupted the ingroup tree. This effect of outgroup inclusion disrupting the ingroup was also found for both mammals and birds [17, 18]. In their

study, Holland et al. [16] used only five taxa; as the number of taxa increases and the models become more complex additional problems are expected.

I.III Materials and Methods

To extend the work of Holland et al. [16], we focused on a symmetric 8-taxon ingroup tree with five short internal branches and a one- or two-taxon outgroup joining at the middle point of the inner-most branch (Figure I). This is a generalized version of a rapid radiation. A symmetric tree was chosen for its potential for analytical (exact) solutions.

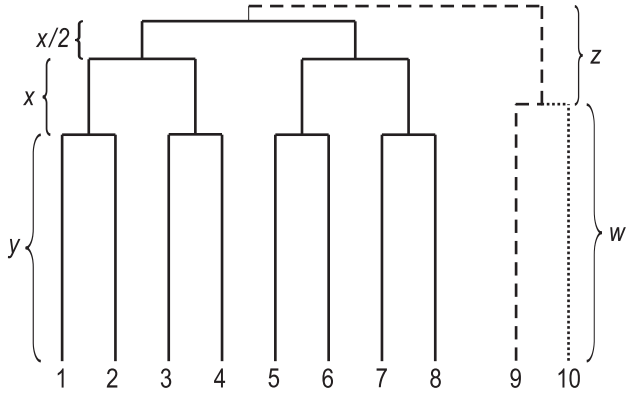


Figure I - The model trees used for simulations (8-taxon simulation: solid lines, 9-taxon simulation (one-taxon outgroup): solid and dashed lines, 10-taxon simulation (two-taxon outgroup): solid, dashed and dotted lines). In the 9-taxon simulation, *z* was set to 0. In both the 9- and the 10-taxon trees the outgroup attaches the ingroup at the middle of the most inner edge of the 8-taxon tree.

For all simulations, unless otherwise stated, the following settings and procedures apply. Seq-Gen version 1.3.2 [19] was used to generate the sequences. Four-state data-sets were generated on each of the trees using the Jukes-Cantor model [20] of nucleotide substitution. We chose to use the Jukes-Cantor model, which is nested within more complex models for 4-state characters [21], to ensure the generality of our results. Substitutions at each site were independent and identically distributed (iid) with equal rates. Branch weights (lengths) were defined to be the expected number of substitutions per site on each branch. Each of the three tree-estimation algorithms maximum parsimony (MP), maximum likelihood (ML) and neighbor-joining (NJ) were applied to every sample sequence using PAUP* version 4b10 [22]. For MP and ML, heuristic searches were done with the HSearch command's default settings except for the option NBest, which was set to 1 (this was done so that, for each dataset, only one best tree discovered during the search will be saved).

When comparing NJ applied to corrected distances and MP, two parameters are being changed simultaneously - the tree building method and whether or not a correction for multiple substitutions is done [23, 24]. However, it is possible to separate the effects of these two parameters to allow for a better comparison between the methods. Therefore, NJ was applied both with the Jukes-Cantor correction [20] by setting the DSet option in PAUP* to JC, and with no correction by setting the DSet option to p. In some cases MP was performed with Jukes-Cantor correction in addition to its usual implementation (no correction). Although correcting MP for multiple changes is possible, it is not implemented in publicly available software. Therefore, correction for MP was implemented using our own code with distance Hadamard [25] applied to distances that were corrected by the Jukes-Cantor method (this code is available from l.shavit@massey.ac.nz). For more information about corrected maximum parsimony and the effect of the correction on parsimony's consistency see Steel, Hendy, and Penny [23] and Penny et al. [24].

Sequences were generated on the model trees depicted in Figure I. Branch lengths varied according to parameters *x*, y, z and w (Figure I), where *x* (ranging from 0.005 to 0.025 in steps of 0.010) is the expected number of substitutions per site on each of the five internal branches, y (ranging from 0.1 to 0.3 in steps of 0.1) is the expected number of substitutions per site on each of the eight external branches, z (ranging from 0 to 0.3 in steps of 0.05) is the expected number of substitutions per site on the edge connecting the outgroup taxa to the ingroup in the middle of the inner-most edge, and w (ranging from 0 to 0.3 in steps of 0.05) is the expected number of substitutions per site on each outgroup branch. If $z+w \geq 1.5x+y$, then there is a point on the tree such that the distances from that point to each of the leaves are all equal (we then say that 'a molecular clock is maintained', though this is not true for all parameter combinations used here). 1000 data-sets were generated of lengths l = (200, 400, 800, and 1600) for each parameter combination of the model tree. The reconstructed unweighted trees (without edge lengths) were compared with the model (generating) tree.

I.IV Results

I.IV.I **8-taxon Simulation**

Accuracy of the methods – We first considered the ability of the methods to reconstruct the ingroup tree alone. Sequences were generated on the 8-taxon tree T8=(((1,2),(3,4)),((5,6),(7,8))) (see Figure I). Figure II shows the accuracy of the different methods in reconstructing T8 for different regions of the parameter space. The results of this simulation show that all four methods are consistent for all regions of the parameter space. As expected, all methods are less accurate when the internal edges are short and the external branches are long.

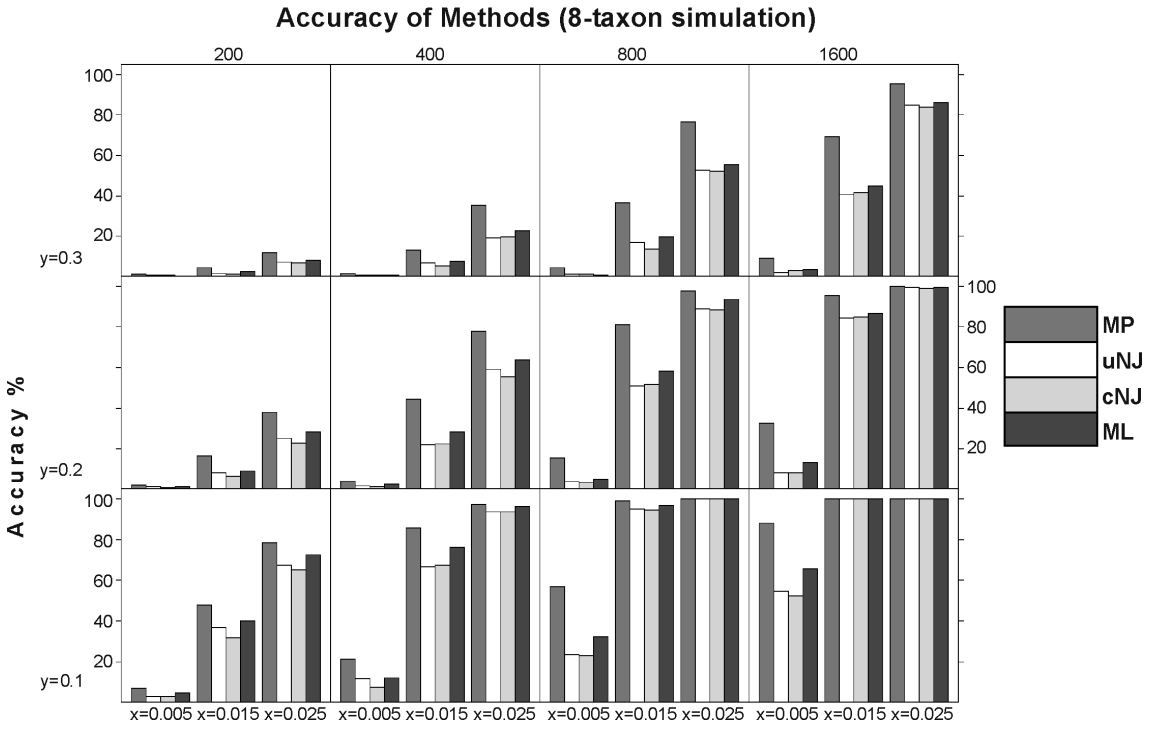


Figure II - Accuracy of maximum parsimony (MP), uncorrected neighbor-joining (uNJ), corrected neighbor-joining (cNJ), and maximum likelihood (ML) in reconstructing the 8-taxon tree. In each box, the percentage of correct trees out of the 1000 trees constructed by each method is shown for each length of the internal edges *x* = (0.005, 0.015, 0.025). Each row corresponds to a different branch length *y* = (0.1, 0.2, 0.3) and each column corresponds to a different sequence length *l* = (200, 400, 800, 1600).

An unexpected feature is that in this parameter space MP performed as well as, and usually better than, the other methods tested. This was surprising because the tree and parameters were chosen so that it would be difficult for MP to obtain the correct tree. However, it is known that some biases can favor the correct tree [26, 27]. In the case of long external branches adjacent to short internal branches, the lengths of the short internal branches are overestimated resulting in the recovery of the correct tree [15, 26, 27, 28, 29, 30, 31]. In our results, the more difficult the parameter combinations were (shorter *x*, and/or longer *y*), the bigger the improvement in accuracy of MP over the other methods. ML performed slightly better than uncorrected NJ (uNJ) and corrected NJ (cNJ). uNJ and cNJ found T8 with virtually the same frequencies, for each point in the parameter space.

Topological Bias – Two trees have the same unlabeled topology if one tree can be converted into the other (ignoring branch lengths) by a permutation of the labels (taxon names). A twofold symmetry is a point on any vertex or edge on the tree where precisely two of the subtrees are topologically identical. An example of a twofold symmetry is a cherry, which is defined as a single pair of leaves adjacent to a common node [32]. Note that a star tree with 3 or more taxa contains no cherries as there are more than two taxa adjacent to the single internal node. We investigated the bias of phylogenetic methods towards estimating trees with a certain number of cherries. The four possible 8-taxon, unrooted, unlabeled, bifurcating tree topologies are shown in Figure III. Their frequencies were calculated using the formula given by Hendy, Little and Penny [33](see also [34]). Within the four possible unlabeled topologies of 8-taxon bifurcating trees, one topology comprises four cherries, two topologies have three cherries and one topology has two cherries (Figure III).

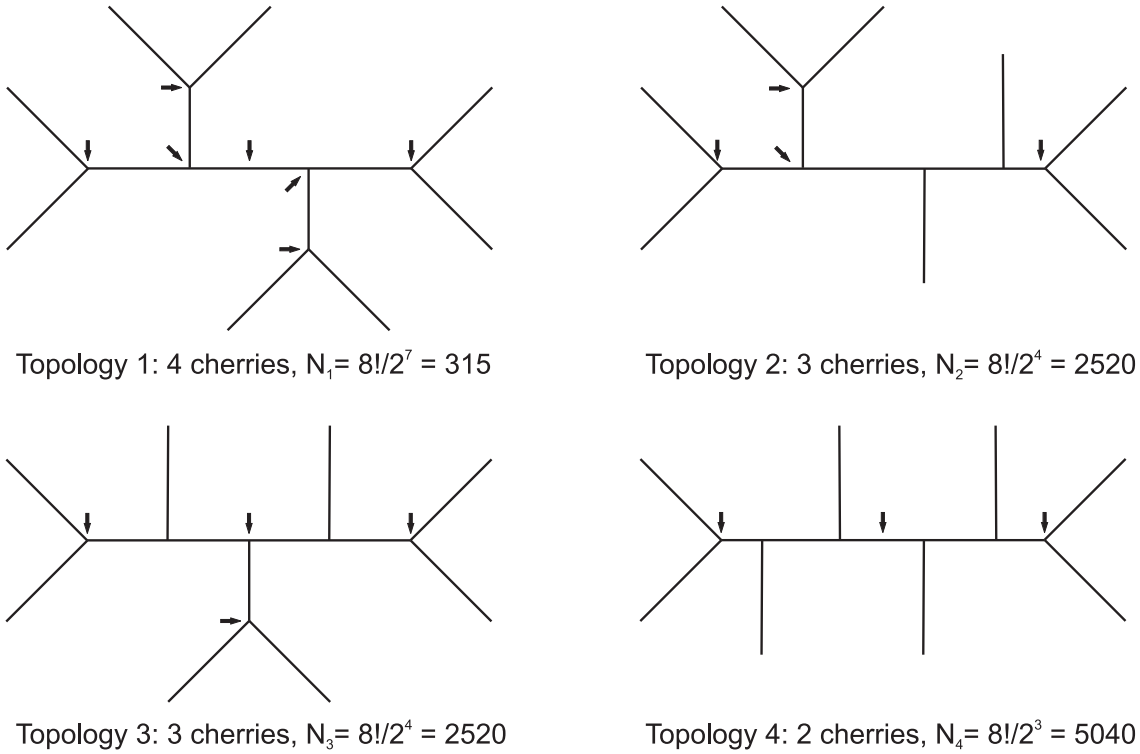


Figure III - The four unlabeled topologies of 8-taxon bifurcating trees. Topology 1 has four cherries, Topology 2 and 3 have three cherries and Topology 4 has two cherries. The twofold centers of symmetry in each topology are indicated by arrows. The number, *Nt,* of different tip-labeled bifurcating trees having each topology t is given.

To test the hypothesis that parsimony methods are biased towards selecting the highly symmetric topology of T8, 10,395 alignments (the number of 8-taxon, unrooted, bifurcating trees) were generated on an 8-taxon star-tree (by setting x=0). The expected number y of substitutions per site on the eight (external) branches was set to 0.2, and the length of the generated sequences was set to 1000. Each of the five phylogenetic methods was applied to the set of alignments, and the number of trees of each of the four topologies (Figure III) was recorded. The DCOLLAPSE and LCOLLAPSE options in PAUP* were both set to 'yes', thus allowing uNJ, cNJ, and ML to collapse branches with length smaller than 10^{-8}. For MP the two COLLAPSE options MINBRLEN and MAXBRLEN were tested.

It is important to note, that since the star-tree is a multifurcating tree with no internal branches, the correct number of trees having any of the four bifurcating topologies is 0. However, seeing that all methods selected many bifurcating trees, we compared the

distribution of these to the distribution of all different 8-taxon, leaf-labeled, unrooted, bifurcating trees. The results are shown in Figure IV. All methods were found to be biased towards fully resolved trees. Even when they were allowed to collapse zero-length branches, none of the methods ever recovered the star-tree. Moreover, the biases demonstrated were not equivalent for all methods.

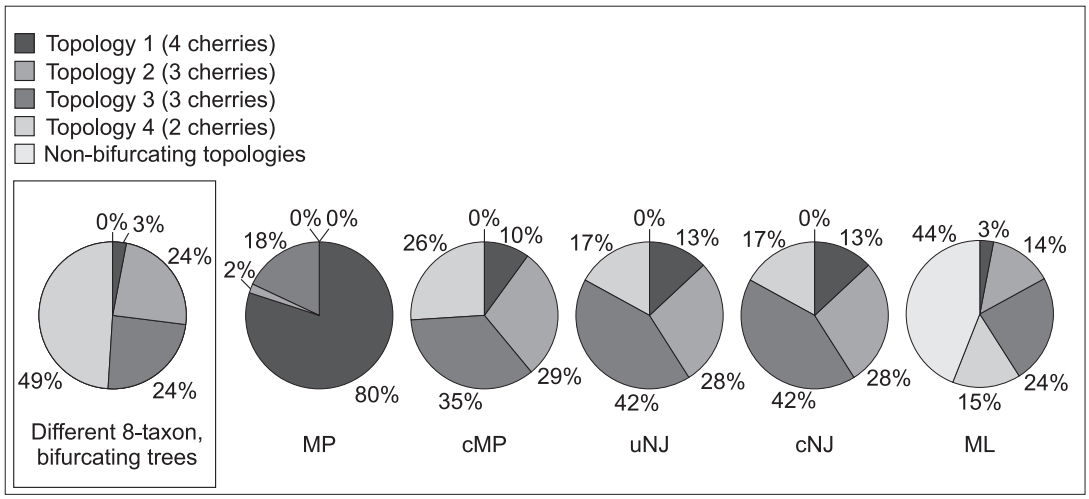


Figure IV - The percentage of trees, having each of the possible topologies for an 8-taxon unrooted tree, out of 10,395 trees constructed by each method for sequences generated on a star tree. The box on the left shows a classification of the 10,395 8-taxon bifurcating trees into the four possible topologies. On the right is the classification of the 10,395 trees, constructed by each method for sequences generated on a star tree, into the four possible unlabeled bifurcating tree topologies.

Strikingly, for MP, 80% of the estimated trees had four cherries, although only 3% (315 out of 10,395) of the 8-taxon bifurcating trees have such a topology. Furthermore, MP did not select any trees with two-cherries or any multifurcating trees. In this example, we did not detect any differences in the results using either of the two collapsing options (MINBRLEN, MAXBRLEN). A less extreme bias was found for corrected MP (cMP), where 10% of the estimated trees had four cherries and 26% had two cherries. Both uNJ and cNJ had similar biases with only 17% of the estimated trees having two cherries, substantially less than the 49% (5,040 out of 10,395) of bifurcating trees having this topology. 13% of the trees constructed by uNJ and cNJ had four cherries, still well in excess of the 3% in the uniform distribution of the bifurcating trees.

Compared with the distribution of bifurcating trees, all methods selected more trees with topology 3, and fewer trees with topology 2 (both topologies have three cherries). For ML, some of the trees with topology 2 were collapsed into multifurcating trees. ML also found fewer trees with topology 4 (two cherries) than there are in the uniform distribution.

For MP, cMP, uNJ and cNJ a general bias towards forming cherries was found. Although ML demonstrated less bias towards forming cherries, it did exhibit bias against collapsing edges that are adjacent to cherries. In more than 40% of the cases, ML selected multifurcating trees; however, the star-tree was never selected. MP, uNJ and cNJ estimated only bifurcating trees, even though the collapse options in PAUP* version 4b10 [22] were set to 'yes'. The bias towards selecting bifurcating trees with cherries is particularly evident for MP, and this is almost certainly the explanation for why MP appears to perform so well in Figure II. When the sequence length was increased to 10,000, cMP, uNJ and cNJ selected each topology with a similar frequency (to within 2%) to that found when the length of the generated sequences was set to 1000. ML selected more trees with two cherries (topology 4) and fewer trees of topology 3 than were selected when the sequence length was set to 1000, but selected the other topologies with similar frequencies (to within 1%) to those found with sequence length l = 1000. MP selected only (i.e. 100%) trees with four cherries (topology 1). We also tested the effect of setting NBEST to 'no', allowing the methods to select more than one tree for each data-set (while weighting the trees for each data-set, so that the total weight of each data-set was 1). This did not have a significant effect on the results.

We have shown that, for the parameter space used, all the phylogenetic methods tested here were consistent in reconstructing the 8-taxon tree (T8). Nonetheless, we found that the phylogenetic methods tested, and particularly MP, are biased towards specific tree topologies.

I.IV.II **Adding a Single-Taxon Outgroup**

The next simulation tested the effect of adding a single-taxon outgroup to the 8-taxon tree. Sequences were generated on the 9-taxon tree T9=((((1,2),(3,4)),((5,6),(7,8))),9). The expected number of substitutions per site on the edge connecting the outgroup taxa to the ingroup, z, was set to 0.

Accuracy of the methods – Given that the simulation study done by Holland et al. [16] found that the addition of an outgroup can disrupt a correct ingroup, we compared the outcomes of applying the methods to the 9-taxon alignment and to an alignment of the eight ingroup taxa alone. The results were classified into six categories according to the scheme shown in Figure Va, based on whether or not the 9-taxon tree (constructed from the 9-taxon alignment) was correct and whether or not the 8-taxon tree (constructed from the 8-taxon ingroup alignment) was correct. The percentage of trials resulting in each category is reported in Figure Vb. As in the 8-taxon simulation, and as expected, all methods were found to be less accurate when the internal edges are short and the external branches are long (see supplementary material 1).

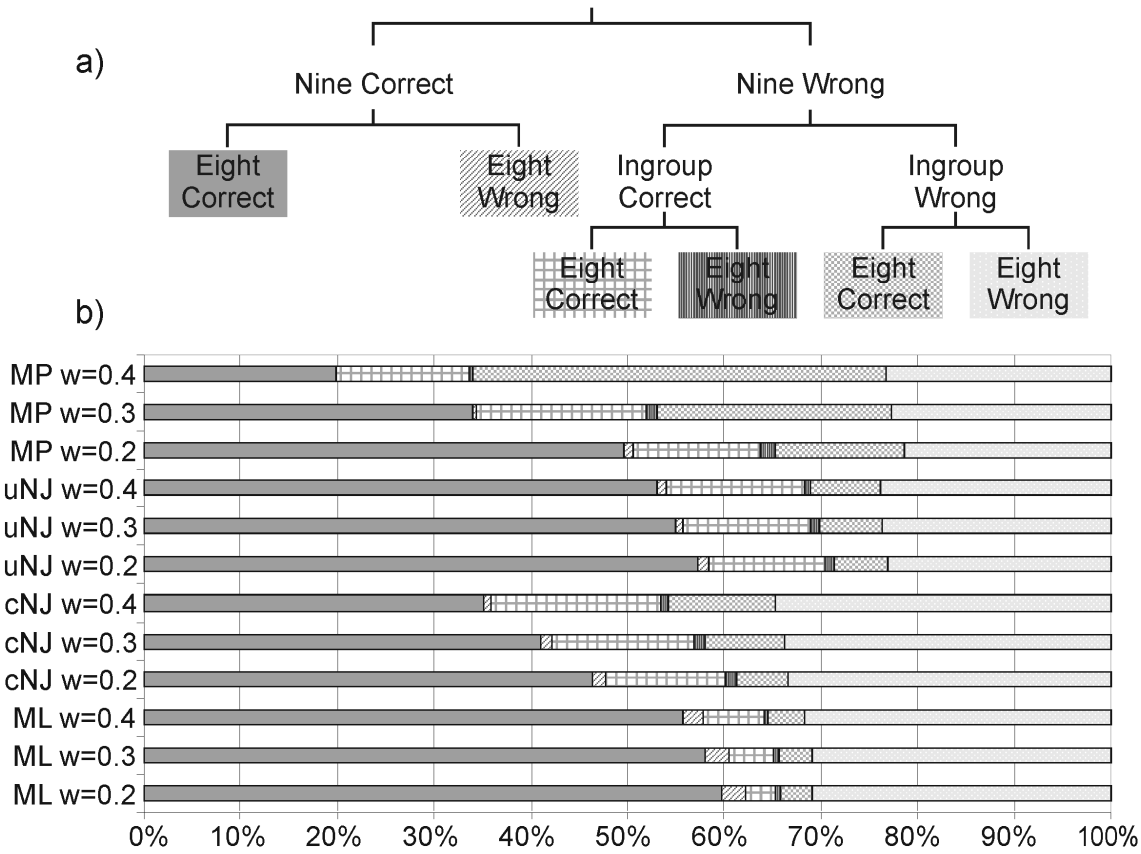


Figure V - Frequencies of different types of error in reconstructing the 9-taxon tree. a) The different types of result combinations in reconstructing the 9- and 8-taxon tree from the 9- and 8-taxon alignments, respectively. Each terminal node description of a category gives the result of 8-taxon estimation, while internal node descriptions are results of 9-taxon estimation. The first category is where both the 8- and 9-taxon trees are correct, the second is where the addition of the outgroup corrects an incorrect 8-taxon tree. The third category is where both the 8-taxon tree and the ingroup tree within the 9-taxon tree were constructed correctly, but the outgroup was misplaced. The fourth category is where the addition of the outgroup corrects an incorrect ingroup tree, but the outgroup itself is placed incorrectly. More disturbing is the fifth category where the inclusion of the outgroup has confounded the correct 8-taxon ingroup tree. The last category is where both the 8-taxon tree and the ingroup within the 9-taxon tree are incorrect. b) The results for sequence length *l* = 1600 averaged over the length of the internal edges, *x*, and the length of the external branches, *y*.

With the inclusion of an outgroup, the accuracy in reconstructing the correct ingroup tree was reduced compared to the 8-taxon case. As expected, the results show that the more distant the outgroup becomes, the more difficult it is to reconstruct the correct tree

(Figure V). For the tree and parameters used, ML was the most accurate of the methods tested. MP, which was very accurate in reconstructing the 8-taxon tree (see topological bias), was particularly affected by the inclusion of the outgroup. In fact, MP was the only method that became inconsistent (with parameters $x = 0.005$ or $x = 0.015$, $y = 0.3$, $z = 0$ and $w = 0.4$; see supplementary material 1). When the molecular clock was maintained uNJ performed better than cNJ, but when the molecular clock was violated cNJ was more accurate (see Table I).

Table I - Accuracy of cNJ and uNJ in reconstructing the 9-taxon tree with respect to the molecular clock assumption.

y	*x*	*w*	Molecular Clock	cNJ	uNJ
0.1	0.005	0.2	YES	143	**177**
0.1	0.005	0.3	YES	102	**159**
0.1	0.005	0.4	YES	58	**138**
0.1	0.015	0.2	YES	863	**893**
0.1	0.015	0.3	YES	771	**859**
0.1	0.015	0.4	YES	657	**838**
0.1	0.025	0.2	YES	974	**978**
0.1	0.025	0.3	YES	946	**973**
0.1	0.025	0.4	YES	896	**972**
0.2	0.005	0.2	NO	**26**	24
0.2	0.005	0.3	YES	10	**14**
0.2	0.005	0.4	YES	6	**7**
0.2	0.015	0.2	NO	**561**	547
0.2	0.015	0.3	YES	418	**491**
0.2	0.015	0.4	YES	340	**449**
0.2	0.025	0.2	NO	**893**	886
0.2	0.025	0.3	YES	807	**848**
0.2	0.025	0.4	YES	736	**834**
0.3	0.005	0.2	NO	**4**	2
0.3	0.005	0.3	NO	**4**	3
0.3	0.005	0.4	YES	2	2
0.3	0.015	0.2	NO	**198**	172
0.3	0.015	0.3	NO	175	175
0.3	0.015	0.4	YES	93	**122**
0.3	0.025	0.2	NO	**635**	575
0.3	0.025	0.3	NO	**568**	564
0.3	0.025	0.4	YES	437	**505**

Results are shown for datasets of 1000 trees and sequence length *l*=1600. Bold font indicates the method with better accuracy out of cNJ and uNJ. Uncorrected NJ is more accurate under the molecular clock assumption, but as this assumption is increasingly violated, correcting for multiple substitutions becomes advantageous.

Most interesting are cases in which the 8-taxon ingroup tree was correct, but adding the outgroup disrupted the ingroup (these are ~13% of all cases). In most of those cases, the distorted ingroup results from the outgroup attaching to the ingroup at one of the long external branches, two branches away from the correct short internal branch. Examples, where the addition of an outgroup distorts an ingroup tree, were previously reported for birds [18] and for mammals [17]. The converse situation, where an incorrect ingroup tree was constructed (on an 8-taxon alignment) but the correct 9-taxon tree was found, occurred in less than 1.5% of the cases (Figure V).

Misleading Zone – For MP, the simulations on the 9-taxon tree with the parameters x=0.015, y=0.2, z=0 and w=0.4 were extended to include sequence lengths of l = (200, 400, 800, …, 204,800). Trees were classified into four categories: 1) the single correct tree; 2) the four trees in which the ingroup phylogeny is correct but the outgroup (taxon 9) is incorrectly joined to one of the internal branches; 3) the eight trees in which the ingroup phylogeny is correct but the outgroup (taxon 9) is incorrectly joined to one of the external branches; 4) the remaining 135,122 trees. The results for sequence lengths l = 200-102,400 are shown in Figure VI.

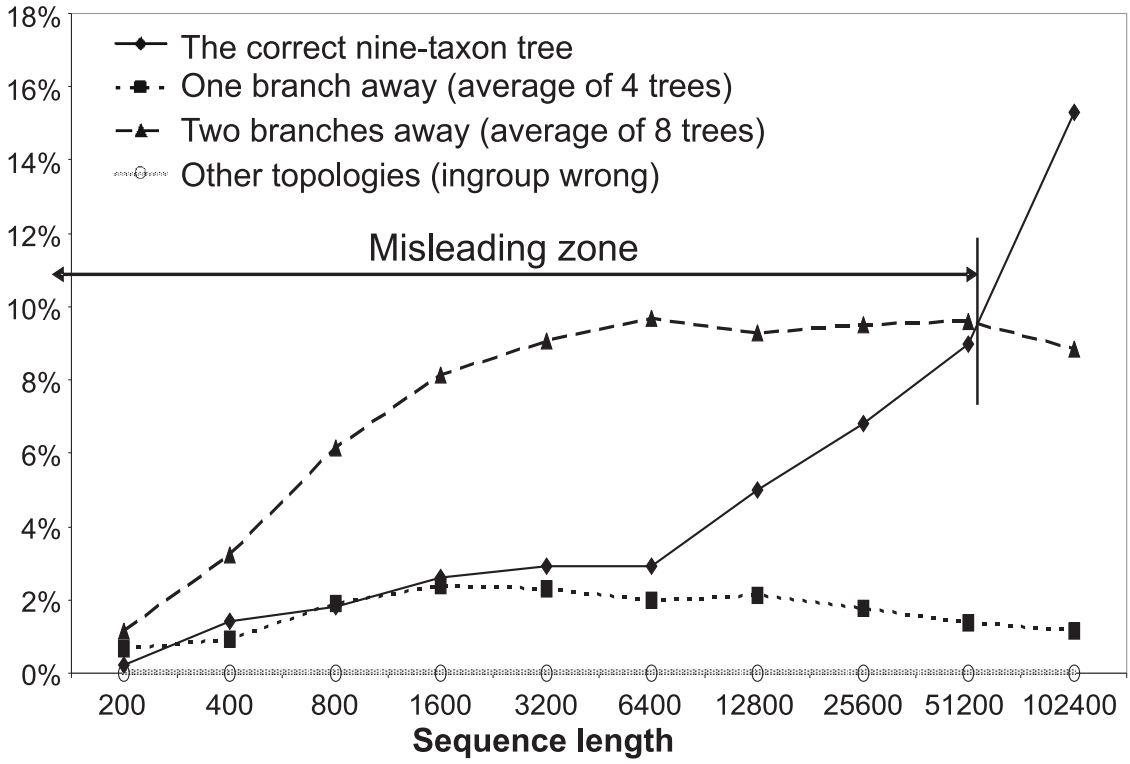


Figure VI - The misleading zone for MP. In this simulation the 9-taxon tree T9=((((1,2),(3,4)),((5,6),(7,8))),9), with parameters *x*=0.015, *y*=0.2, *z*=0, *w*=0.4, was used. The frequency with which the correct tree and each of the competing trees - the four (one branch away) trees in which the ingroup phylogeny is correct but the outgroup, taxon 9, is incorrectly joined to one of the internal branches (for example ((((1,2),9),(3,4)),((5,6),(7,8)))), and the eight (two branches away) trees in which the ingroup phylogeny is correct but the outgroup, taxon 9, is incorrectly joined to one of the external branches (for example ((((1,9),2),(3,4)),((5,6),(7,8)))) - were chosen is shown. All other 135,122 trees in which the ingroup is wrong are collectively referred to as "other topologies". For each category the results are averaged over the number of trees in the category. The misleading zone extends to a sequence length of approximately 60,000 nucleotides. Only then does the correct tree get selected more frequently than each of the eight competing trees that are two branches away from the correct tree.

Within its consistency zone, the probability of MP selecting the correct tree goes to 1 as the sequence length increases. However, following Holland et al. [16] we have identified a misleading zone within, but close to the boundary of, the consistency zone of MP. This is a specific region of the parameter space in which MP is consistent, but for finite sequence lengths it is possible for each of several individual incorrect trees to be selected more frequently than the correct tree. For example, the 9-taxon tree with

parameters x=0.015, y=0.2, z=0 and w=0.4 is inside the misleading zone of MP. For l=1600, each of 8 incorrect trees is selected with much greater frequency than the correct tree. For l=200 (using 10,000 data-sets), we found a ratio of ~1:3 between the correct tree and each of eight incorrect trees where the outgroup attaches to one of the external branches. Sequences of ~60,000 nucleotides are required before the correct tree is chosen more frequently than any other tree. With sequence length of 102,400 the correct tree is still only recovered ~15% of the time. With sequence length of 204,800 (not shown) the correct tree is recovered in ~28% of the time. Extrapolating from this data, we expect that a sequence length of at least 400,000 characters would be needed for MP to have a 50% success rate in finding the correct tree. It is important to note, that correcting for multiple substitutions significantly reduces the size of maximum parsimony's misleading zone for this combination of parameters. In fact, for cMP (as for uNJ, cNJ and ML), a sequence length as short as 400 is already enough for the correct tree to be chosen most frequently (data not shown). For short sequence lengths (l=200), all methods often select an incorrect tree and some incorrect trees are each selected with greater frequency than that of the correct tree. But since the number of times each tree is selected is very small, it is difficult to check whether this is statistically significant. However, as in the 5-taxon study of Holland et al. [16], there does appear to be a small misleading zone for all the methods studied here.

Breaking symmetry – In order to evaluate the effect of breaking the symmetry of the ingroup tree, we changed the 9-taxon tree so that one external edge of the ingroup is longer than the others (a higher rate of evolution), and consequentially the symmetry of the ingroup is broken. The results (Figure VII) show that the longer this ingroup branch is, the more frequently the outgroup joins it, reducing the accuracy of all methods in reconstructing the 9-taxon tree. While the long external edge seems to have little effect on the accuracy of ML and cNJ, a strong negative effect on both MP and uNJ was observed. The longer the selected external edge, the further we are from maintaining a molecular clock, and the more pronounced the advantage of the corrected methods (ML and cNJ) over the uncorrected methods (uNJ and MP) becomes.

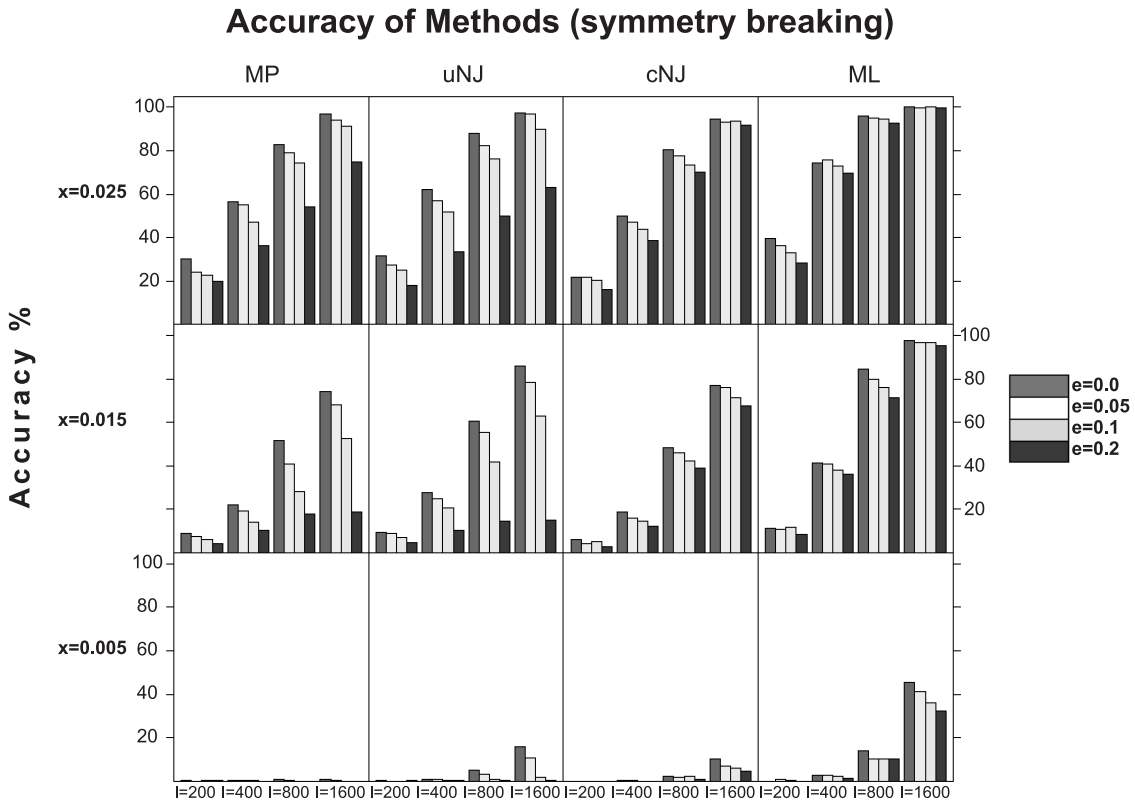


Figure VII - The effect of symmetry breaking. The results are shown for parameters *y*=0.1 and *w*=0.3. In each box, the percentage of correct trees reconstructed for each sequence length and each of the four lengths *e*=(0, 0.05, 0.1, 0.2) that were added to one external branch is shown. Each row corresponds to a different length of internal edges *x*=(0.005, 0.015, 0.025) and each column corresponds to a different tree-estimation method.

We have demonstrated that with the use of a single-taxon outgroup and a rapid radiation, it is difficult to correctly infer the position of the root, even when the ingroup tree is correct. This is particularly noticeable when the substitution rate of one ingroup taxon is higher than the others. Of particular concern is the observation that introducing an outgroup can interfere with the accuracy of the ingroup tree.

I.IV.III **Two-Taxon Outgroup**

Accuracy of the methods – This simulation was used to evaluate the effect of including a second outgroup-taxon, on the accuracy of the different methods in reconstructing the tree. Sequences were generated on the 10-taxon tree T10 (Figure I) with two, one or zero outgroup taxa removed to acquire the 8-, 9- and 10-taxon datasets,

respectively. The phylogenetic methods were applied to the same data-sets, and their ability to reconstruct the correct 8-, 9- and 10-taxon trees was compared.

In Table II, the number of times in which the tree was reconstructed correctly is reported for each of the methods and for the four different branch lengths used. In every single case, correct trees were reconstructed more frequently for the 10-taxon data compared to the 9-taxon data. However, the frequency with which trees were correctly estimated for the 8-taxon data is higher than for both the 9-taxon and 10-taxon data-sets. This is true for all four methods with each of the sequence lengths.

Table II - Accuracy of the final tree.

Method	Length	8 Correct	9 Correct	10 Correct
MP	200	156	13	28
MP	400	457	40	117
MP	800	804	92	302
MP	1600	971	221	573
uNJ	200	75	11	16
uNJ	400	204	48	67
uNJ	800	506	179	222
uNJ	1600	837	493	562
cNJ	200	71	7	12
cNJ	400	197	37	52
cNJ	800	499	144	188
cNJ	1600	829	430	513
ML	200	86	19	24
ML	400	249	109	124
ML	800	555	383	423
ML	1600	857	777	815

Accuracy of the methods in reconstructing the 8- 9- and 10-taxon trees from the respective sequence data, with the parameters set to x=0.015, y=0.2 and z+w=0.3. The results are averaged over the five central values of z = (0.05, 0.1, 0.15, 0.2, 0.25). Correct trees were reconstructed more frequently for the 8-taxon data than for the 9- and 10-taxon sequence data. In every case, more correct trees were constructed for the 10-taxon data than for the 9-taxon data.

This increase in reliability, when going from 9 to 10 taxa, runs counter to our intuition that the greater the number of taxa (and so the greater the number of internal edges that need to be estimated) the more difficult it is to reconstruct the correct tree. A possible explanation is that the more balanced topology of the 10-taxon tree makes it easier for the methods to reconstruct it. The correct ingroup is reconstructed most frequently for the 8-taxon (ingroup alone) data-sets (Table III), and more frequently for the 10-taxon data-set than for the 9-taxon data-sets. Thus, the inclusion of a single-taxon outgroup disrupts the correctly constructed ingroup more frequently than does the inclusion of the two related outgroup taxa.

Table III - Ingroup tree accuracy.

Method	Length	8-taxon Correct	9-taxon Ingroup Correct	10-taxon Ingroup Correct
MP	200	156	56	80
MP	400	457	145	271
MP	800	804	328	593
MP	1600	971	558	877
uNJ	200	75	44	56
uNJ	400	204	137	163
uNJ	800	506	409	457
uNJ	1600	837	790	818
cNJ	200	71	35	46
cNJ	400	197	113	140
cNJ	800	499	369	419
cNJ	1600	829	751	792
ML	200	86	60	64
ML	400	249	201	211
ML	800	555	522	540
ML	1600	857	860	868

Accuracy of the methods in reconstructing the ingroup tree for the 8- 9- and 10-taxon sequence data with the parameters set to x=0.015, y=0.2 and z+w=0.3. The results are averaged over the five central values of z = (0.05, 0.1, 0.15, 0.2, 0.25). Ingroup correct trees were reconstructed more frequently for the 8-taxon data than for the 9- and 10-taxon sequence data. The addition of a one-taxon outgroup disrupts the ingroup tree more frequently than the addition of a two-taxon outgroup.

Placement of a second outgroup-taxon – Biologists often face the problem of choosing good outgroup taxa for tree reconstruction. In this simulation we tested how the placement of the second outgroup taxon affects the accuracy of the methods in reconstructing the ingroup, i.e. the 8-taxon, tree. The ability of the methods to reconstruct the ingroup for different values of z (the expected number of substitutions on the edge connecting the outgroups' common ancestor to the ingroup) and *w* (the expected number of substitutions on the edge of each outgroup-taxon) was compared. In addition, the outcomes of these runs were compared with the corresponding results for nine and eight taxa (all phylogenetic methods used were applied to the same data-sets). The results are shown in Figure VIII, where the accuracy of the methods in reconstructing the ingroup tree using the eight, nine and ten taxa (unconstrained) is presented. The results are categorized into eight categories: 'rrr') ingroup correct in all (8-, 9- and 10-taxon); 'rrw') ingroup wrong in the 10-taxon but correct in the 8- and 9-taxon; 'rwr') ingroup correct in the 8- and 10-taxon but wrong in the 9-taxon; 'rww') ingroup wrong in both the 9- and 10-taxon but correct in the 8-taxon; 'wrr') ingroup correct in the 9- and 10-taxon but wrong in the 8-taxon; 'wrw') ingroup wrong in the 8- and 10-taxon but correct in the 9-taxon; 'wwr') ingroup correct in the 10-taxon but wrong in the 8- and 9-taxon; 'www') ingroup wrong in all (8-, 9- and 10-taxon).

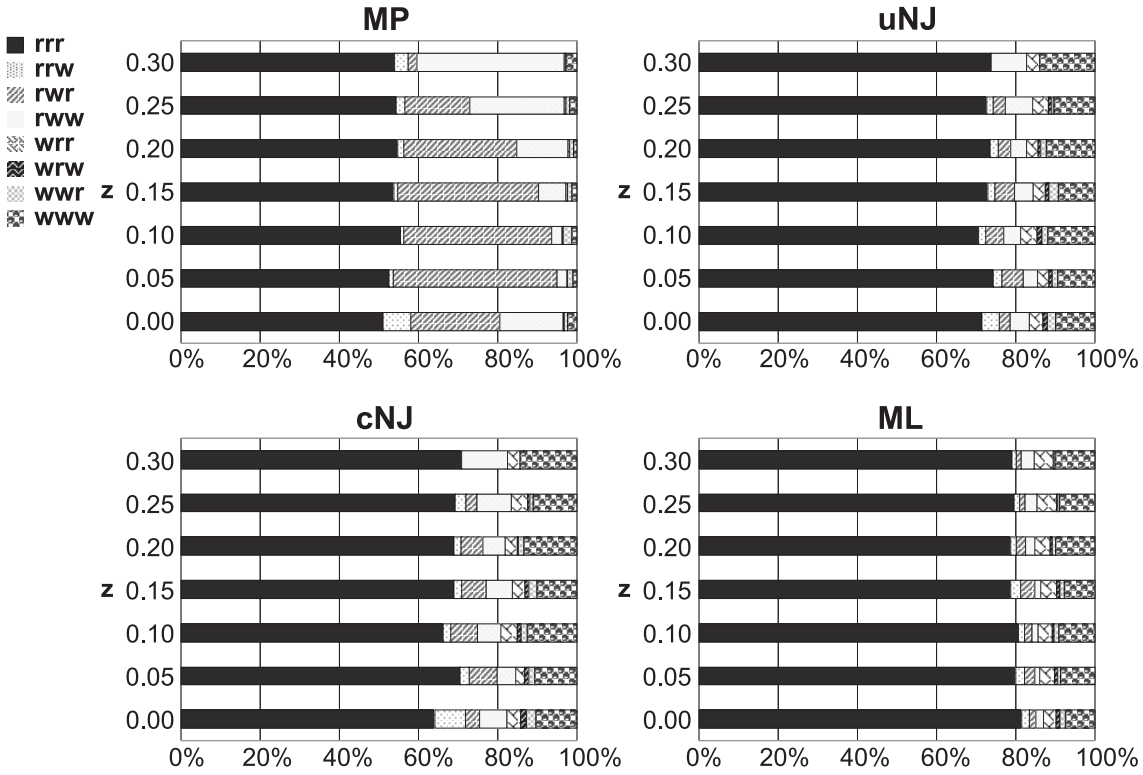


Figure VIII - Classification of the 10-taxon results into the eight possible combinations (see text). In this simulation a 10-taxon tree with parameters *x*=0.015, *y*=0.2 and *z*+*w*=0.3 was used, *z* varied from 0 to 0.3 in steps of 0.05. For each method the percentage of trees out of 1000 trees constructed, in each class described in the text, is shown for each length of the edge connecting the outgroup taxa to ingroup tree. The results are shown for sequence length *l*=1600.

As expected, ML constructed the correct ingroup (for 8-, 9- and 10-taxa) more frequently than did the other methods. Although uNJ performed slightly better than cNJ, both constructed the correct trees with similar frequencies (the parameters used obey the molecular clock assumption). MP reconstructed the ingroup correctly for all in only about 55% of the cases; however it had the lowest percentage of ‘www‘ (wrong in all). Moreover, when the common ancestor of the two outgroup taxa was close to the ingroup (*z*=0.05), MP reconstructed the ingroup tree correctly for the 8- and 10-taxon data approximately 95% of the time. In addition, MP has the highest percentage of runs in which the ingroup was reconstructed correctly in the 8- and 10-taxon, but was wrong in the 9-taxon data (‘rwr’), and the lowest percentage of runs in which the ingroup was wrong for the 8-taxon data but was right for the others (‘wrr’). These results are as expected, taking into account the bias parsimony has towards forming cherries. Cases in

which the methods construct the ingroup incorrectly from the 8- and 10-taxon data-sets while reconstructing the correct ingroup from the 9-taxon data-set are very rare (<2%).

Finally, we tested the accuracy of the phylogenetic methods in reconstructing the 10-taxon tree for different numbers (z) of substitutions per site on the edge connecting the common ancestor of the two outgroup taxa to the ingroup, and the effect of constraining the two outgroup taxa to be together. The results are shown in Figure IX. The closer the common ancestor of the two outgroup taxa was to the ingroup (the further the two outgroup taxa are from each other), the more accurate the methods were in reconstructing the 10-taxon generating tree. However, it appears advantageous for z to be larger than 0 (such that there is a split separating the outgroup taxa from the ingroup). This trend is very obvious for MP, where the accuracy dropped very rapidly as the common ancestor of the two outgroup taxa became further from the ingroup. This trend is also noticeable for uNJ and cNJ where a more moderate change in accuracy was observed. For ML, although only a very slight drop in accuracy was found, the general trend still applies. We also found that for z=0 constraining the two outgroup taxa to come together had a positive effect on the accuracy of all the methods, both in reconstructing the ingroup tree and in placing the outgroup taxa in the correct position. When z>0, for long sequences, constraining the two outgroup taxa to come together did not effect the accuracy with which the methods reconstructed the ingroup tree and placed the outgroup taxa (Figure IXa). However, for short sequences and small values of z, a slight improvement was recorded (Figure IXb).

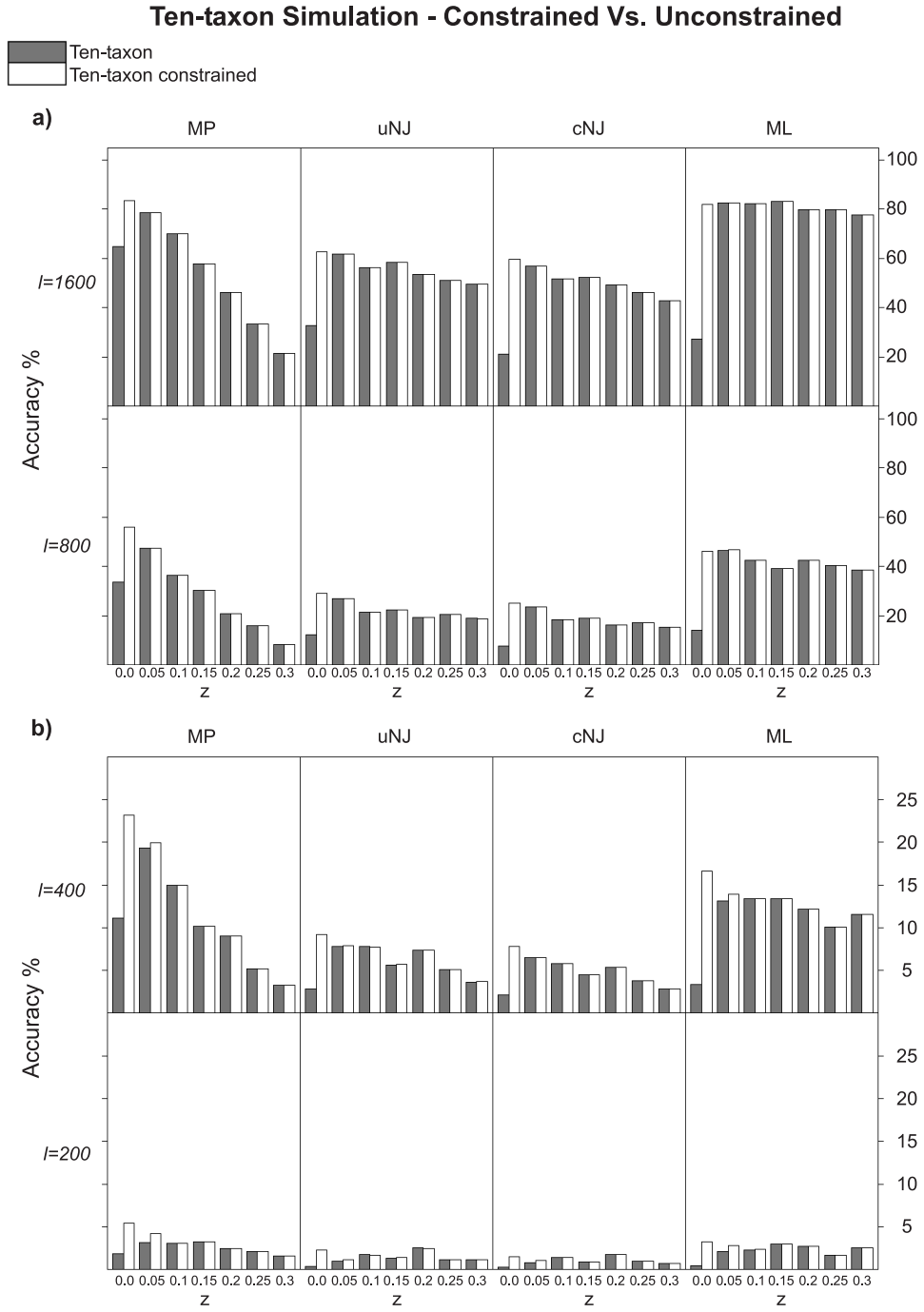


Figure IX - Accuracy in constructing the 10-taxon tree for different number (*z*) of substitutions per site on the edge connecting the common ancestor of the two outgroup taxa to the ingroup, with and without constraining the two outgroup taxa to be together. A 10-taxon tree with parameters *x*=0.015, *y*=0.2 and *z*+*w*=0.3 was used, *z* varied from 0 to 0.3 in steps of 0.05. a) the results for sequence length *l*=(200, 400); scale=(0, 25). b) the results for sequence length *l*=(800, 1600); scale=(0, 100). All methods are more accurate when the two-outgroup taxa used are separated from the ingroup with a common non-zero branch and when the common ancestor of the two outgroup taxa is close to the ingroup.

I.V Discussion

In this simulation study, we have identified problems that are likely to affect the ability of phylogenetic methods to reconstruct tree topologies corresponding to rapid radiations (where there is a combination of short internal and long external branches). Rapid radiations are often star-like, and it is therefore important to identify possible biases in reconstructing a star tree. We established that MP, cMP, uNJ and cNJ are all biased towards forming cherries (see Figure IV). This effect is most pronounced for MP, for which trees having four cherries were chosen many more times than any other topology even though the generating tree had no cherries. ML seems to be biased in a different way; it appears to collapse edges that are not adjacent to cherries. All methods are biased towards a high number of internal edges as none of the methods was successful in recovering the star-tree, even when collapsing was allowed. This effect is similar to the Bayesian "star paradox", where sequences that have evolved on a star-tree can give branches with posterior probability close to one. Steel and Matsen [35] showed that for Bayesian analysis this effect is not expected to automatically vanish given long enough sequences. Topological biases, such as the bias towards forming cherries found here, may work either against or in favor of the methods in reconstructing trees (depending on the true topology of the tree).

Our findings indicate that rooting a star-like tree (many short internal branches connecting long external branches), by joining distant outgroup taxa to a short internal edge, often prevents the correct construction of the ingroup tree (see Table III and Figure V). The effect is particularly strong when an outgroup-taxon and an ingroup taxon share a higher substitution rate (Figure VII). In many of the cases tested, the outgroup was placed two branches away from the correct position. For our data, an important finding is that when a tree rooted by an outgroup is in disagreement with the unrooted ingroup tree, the unrooted ingroup tree is most often correct. For the cases tested here, we found that the use of two outgroup taxa is better than the use of a single outgroup-taxon, both for the accuracy with which a tree is rooted and for maintaining the correct ingroup tree (see Tables 2 and 3). However, ingroup tree reconstruction is

more accurate when the methods are applied to the ingroup alone (see Table III and Figure VIII). We also found that using two outgroup taxa that are distant from each other is better than using two closely related outgroup taxa; this is especially true for MP. For the trees and parameters tested here, and for short sequence lengths, constraining the two outgroup taxa to come together is generally advantageous, especially when they are not closely related (see Figure IXb). However, for longer sequences, constraining the outgroup taxa to come together does not have an effect on the accuracy of the methods (see Figure IXa). In general our results confirm that it is "best practice" to infer phylogeny both with, and without, an outgroup and then compare the results.

Correcting MP for multiple changes was found to be beneficial in the cases where the molecular clock assumption is valid, particularly in cases where MP is misleading or inconsistent. A possible explanation for this is that with the given tree topology under the molecular clock, MP suffers from long-branch attraction. With our parameters, cMP does not suffer from long-branch attraction and therefore is doing better in estimating the correct tree. Nevertheless, under the set of parameters used here, when the molecular clock assumption is violated, MP does not suffer from the long-branch attraction and is indeed biased towards the correct tree. Consequently, under our conditions when the molecular clock assumption is violated, MP is more accurate than cMP in reconstructing the correct tree. This effect is likely to be a characteristic of the highly symmetric model tree.

In the cases where the molecular clock assumption is valid, uNJ was found to be more accurate than cNJ in reconstructing both the 9-taxon tree as a whole and the relationships amongst the ingroup-taxa (see Table). However, when this assumption is violated, by breaking the symmetry of the tree, cNJ and ML were found to be more accurate than uNJ and MP. This effect under the molecular clock may be due to amplification of sampling error and/or because the standard correction has a bias toward overcorrecting. These results are consistent with those found in other simulation studies [16, 36, 37], where corrections for multiple substitutions were found to be helpful only for recovering trees with unequal rates of change along branches. Nei and Kumar [38]

offered guidelines for constructing phylogenetic trees; our results support their argument that uncorrected distances give the correct tree more often than corrected distances when the rate of nucleotide substitution is nearly the same for all evolutionary lineages and there is no strong transition/transversion bias.

Our results support the observation of Holland et al. [16] that methods can be consistent but misleading (even in the absence of model misspecification). We observed a misleading zone for MP, where although the frequency with which the correct tree is found tends to one as the sequence length *l* tends to infinity, for finite yet very long sequences, a number of incorrect trees are each chosen more frequently than the correct tree.

Holland et al. [16] considered the boundary of the consistency zone for MP, i.e. the part of the parameter space where a slight change in the edge lengths makes parsimony either consistent or inconsistent. For 5-taxon trees with 2-state data they calculated that each of the four incorrect trees where the outgroup is drawn to an external edge is selected by MP twice as frequently as the correct tree. In the 5-taxon case, there are only six splits for which the number of substitutions needed is not the same in the five competing trees. All six splits have the same expected frequency on the boundary of MP inconsistency. Two of those splits support the correct tree and each of these has to independently compete with two of the other splits (see [16]). The calculation for four state-data is more complex, but we suspect that the ratio between the correct tree and each of the frequently selected incorrect trees will be equivalent to that of the two-state data. The calculation for the 9-taxon tree is more difficult, as there are many inter-dependent splits. Therefore, further mathematical work is required to calculate the ratio between the correct tree and each of the incorrect trees where the outgroup is drawn to an external edge, and to evaluate the effect of the number of taxa on the frequency with which the correct tree (with the outgroup in its correct placement) is found.

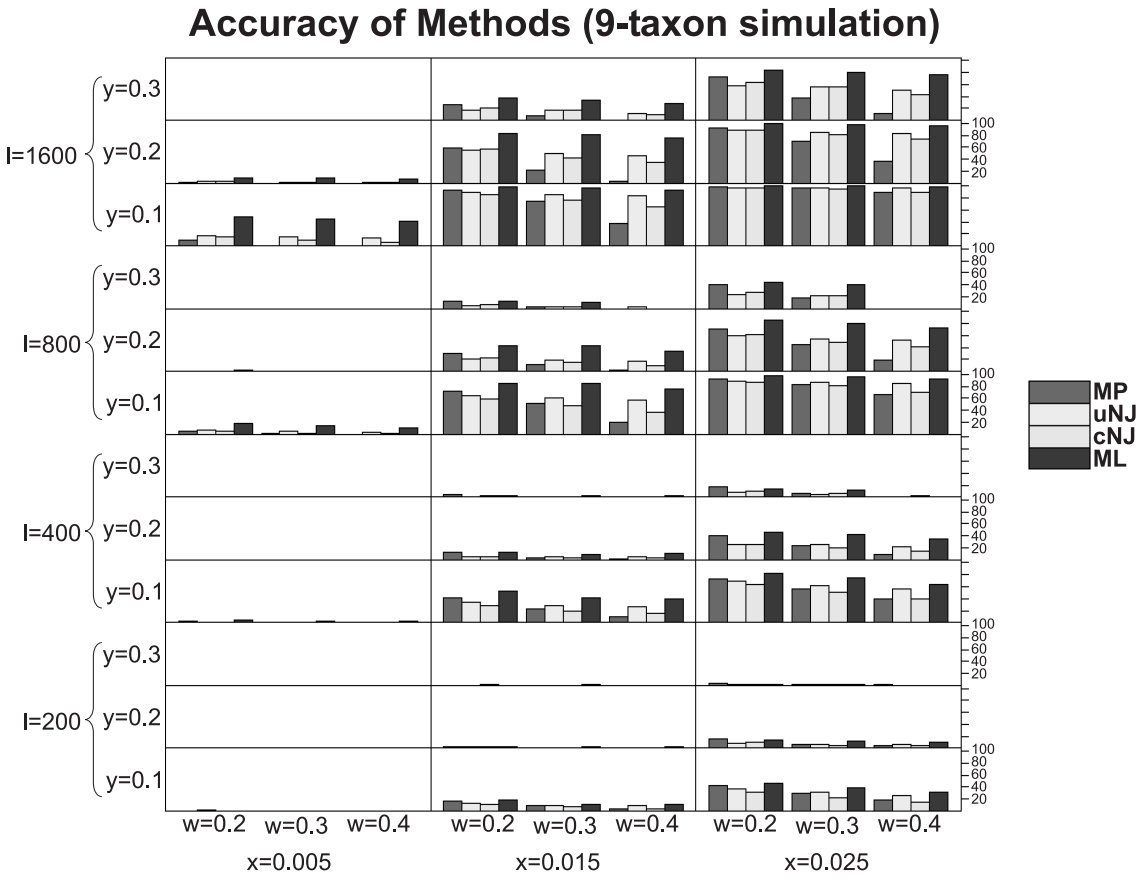
Although this study specifically tested the effects on the reconstruction of an 8-taxon symmetric tree and a simple (biologically oversimplified) substitution model, the problems reported are expected to exist in larger trees and with more complex models

(in which the Jukes-Cantor model is nested). Using a complex model of sequence evolution would not have ensured that any tree estimation properties found were general. In our study, we used four-state data, which is the natural biological language and is known to saturate slightly slower than two-state data [39]. It would be interesting to test the methods further using twenty-state amino-acid data. Bayesian phylogenetic analysis was suggested to be as robust to relative branch-length differences as ML [40], therefore it would also be interesting to test Bayesian inference for the cases studied here.

I.VI Acknowledgements

We thank Klaus Schliep for R-code to generate graphs and statistical advice, and we thank Warwick Allen for computer support. We also thank the Marsden Fund and FRST for funding. This study would not have been possible without the use of Helix parallel computing facility (http://helix.massey.ac.nz).

I.VII Supplementary Material



Accuracy of methods in reconstructing the 9-taxon tree. In each box, the percentage of correct trees out of 1000 trees that were constructed by each method is shown for each length of the outgroup branch *w*=(0.2, 0.3, 0.4). Each row corresponds to a different external branch length *y*=(0.1, 0.2, 0.3) and each column corresponds to a different internal branch length *x*=(0.005, 0.015, 0.025). The results are shown for sequences lengths *l*=(200, 400, 800, 1600). All methods were found to be less accurate when the internal branches were short and the external branches were long.

I.VIII Literature cited

1. Felsenstein J. 1978. Cases in Which Parsimony or Compatibility Methods Will Be Positively Misleading. Syst Zool 27:401-410.
2. Hendy MD and Penny D. 1989. A Framework for the Quantitative Study of Evolutionary Trees. Syst Zool 38:297-309.
3. Bergsten J. 2005. A review of long-branch attraction. Cladistics 21:163-193.
4. Harrison GL, McLenachan PA, Phillips MJ, Slack KE, Cooper A and Penny D. 2004. Four new avian mitochondrial genomes help get to basic evolutionary questions in the late Cretaceous. Mol Biol Evol 21:974-983.
5. Lockhart PJ and Cameron SA. 2001. Trees for bees. Trends Ecol Evol 16:84-88.
6. Lin YH, McLenachan PA, Gore AR, Phillips MJ, Ota R, Hendy MD and Penny D. 2002. Four new mitochondrial genomes and the increased stability of evolutionary trees of mammals from improved taxon sampling. Mol Biol Evol 19:2060-2070.
7. Philip GK, Creevey CJ and McInerney JO. 2005. The Opisthokonta and the Ecdysozoa may not be clades: Stronger support for the grouping of plant and animal than for animal and fungi and stronger support for the Coelomata than Ecdysozoa. Mol Biol Evol 22:1175-1184.
8. Philippe H, Lartillot N and Brinkmann H. 2005. Multigene analyses of bilaterian animals corroborate the monophyly of Ecdysozoa, Lophotrochozoa, and Protostomia. Mol Biol Evol 22:1246-1253.
9. Soltis DE, Albert VA, Savolainen V, Hilu K, Qiu YL, Chase MW, Farris JS, Stefanovic S, Rice DW, Palmer JD and Soltis PS. 2004. Genome-scale data, angiosperm relationships, and 'ending incongruence': a cautionary tale in phylogenetics. Trends Plant Sci 9:477-483.
10. Stefanovic S, Rice DW and Palmer JD. 2004. Long branch attraction, taxon sampling, and the earliest angiosperms: Amborella or monocots? BMC Evol Biol 4.

11. Goremykin VV, Holland B, Hirsch-Ernst KI and Hellwig FH. 2005. Analysis of Acorus calamus chloroplast genome and its phylogenetic implications. Mol Biol Evol 22:1813-1822.

12. Leebens-Mack J, Raubeson LA, Cui LY, Kuehl JV, Fourcade MH, Chumley TW, Boore JL, Jansen RK and dePamphilis CW. 2005. Identifying the basal angiosperm node in chloroplast genome phylogenies: Sampling one's way out of the felsenstein zone. Mol Biol Evol 22:1948-1963.

13. Lockhart PJ and Penny D. 2005. The place of Amborella within the radiation of angiosperms. Trends Plant Sci 10:201-202.

14. Poe S and Swofford DL. 1999. Taxon sampling revisited. Nature 398:299-300.

15. Ho SY and Jermiin *L*. 2004. Tracing the Decay of the Historical Signal in Biological Sequence Data. Syst Biol 53:623-637.

16. Holland BR, Penny D and Hendy MD. 2003. Outgroup misplacement and phylogenetic inaccuracy under a molecular clock - A simulation study. Syst Biol 52:229-238.

17. Lin YH, Waddell PJ and Penny D. 2002. Pika and vole mitochondrial genomes increase support for both rodent monophyly and glires. Gene 294:119-129.

18. Slack KE, Janke A, Penny D and Arnason U. 2003. Two new avian mitochondrial genomes (penguin and goose) and a summary of bird and reptile mitogenomic features. Gene 302:43-52.

19. Rambaut A and Grassly NC. 1997. Seq-Gen: An application for the Monte Carlo simulation of DNA sequence evolution along phylogenetic frees. Computer Applications in the Biosciences 13:235-238.

20. Jukes TH and Cantor CR. 1969. Evolution of protein sequences. Pp. 21-123 in Munro HN, ed. Mammalian protein metabolism. Academic Press, New York.

21. Felsenstein J. 2004. Inferring phylogenies. Sinauer Associates, Inc., Sunderland, Massachusetts.

22. Swofford DL. 2002. PAUP*. Phylogenetic Analysis Using Parsimony (*and Other Methods). Sinauer Associates, Sunderland, Massachusetts.

23. Steel MA, Hendy MD and Penny D. 1993. Parsimony Can Be Consistent. Syst Biol 42:581-587.

24. Penny D, Hendy MD, Lockhart PJ and Steel MA. 1996. Corrected parsimony, minimum evolution, and hadamard conjugations. Syst Biol 45:596-606.

25. Hendy MD and Penny D. 1993. Spectral-Analysis of Phylogenetic Data. J Classif 10:5-24.

26. Swofford DL, Waddell PJ, Huelsenbeck JP, Foster PG, Lewis PO and Rogers JS. 2001. Bias in phylogenetic estimation and its relevance to the choice between parsimony and likelihood methods. Syst Biol 50:525-539.

27. Sullivan J and Joyce P. 2005. Model selection in phylogenetics. Annu Rev Ecol Evol Syst 36:445-466.

28. Yang ZH. 1997. How often do wrong models produce better phylogenies? Mol Biol Evol 14:105-108.

29. Siddall ME. 1998. Success of parsimony in the four-taxon case: Long-branch repulsion by likelihood in the Farris Zone. Cladistics-the International Journal of the Willi Hennig Society 14:209-220.

30. Bruno WJ and Halpern AL. 1999. Topological bias and inconsistency of maximum likelihood using wrong models. Mol Biol Evol 16:564-566.

31. Sullivan J and Swofford DL. 2001. Should we use model-based methods for phylogenetic inference when we know that assumptions about among-site rate variation and nucleotide substitution pattern are violated? Syst Biol 50:723-729.

32. McKenzie A and Steel M. 2000. Distributions of cherries for two models of trees. Math Biosci 164:81-92.

33. Hendy MD, Little CHC and Penny D. 1984. Comparing Trees with Pendant Vertices Labeled. Siam Journal on Applied Mathematics 44:1054-1065.

34. Penny D, Hendy MD and Steel MA. 1991. Testing the Theory of Descent. Pp. 155-183 in Miyamoto MM and Cracraft J, eds. Phylogenetic Analysis of DNA sequences. Oxford university press, New York.

35. Steel M and Matsen FA. 2007. The Bayesian "star paradox" persists for long finite sequences. Mol Biol Evol 24:1075-1079.

36. Saitou N and Imanishi T. 1989. Relative Efficiencies of the Fitch-Margoliash, Maximum-Parsimony, Maximum-Likelihood, Minimum-Evolution, and Neighbor-Joining Methods of Phylogenetic Tree Construction in Obtaining the Correct Tree. Mol Biol Evol 6:514-525.

37. Sourdis J and Krimbas C. 1987. Accuracy of Phylogenetic Trees Estimated from DNA-Sequence Data. Mol Biol Evol 4:159-166.

38. Nei M and Kumar S. 2000. Molecular Evolution and Phylogenetics. OXFORD University Press, Inc., New York.

39. Penny D, McComish BJ, Charleston MA and Hendy MD. 2001. Mathematical elegance with biochemical realism: The covarion model of molecular evolution. J Mol Evol 53:711-723.

40. Mar JC, Harlow TJ and Ragan MA. 2005. Bayesian and maximum likelihood phylogenetic analyses of protein sequence data under relative branch-length differences and model violation. BMC Evol Biol 5.