The dissertation committee for Mark Travis Holder certifies that this is the approved version of the following dissertation:

Using a Complex Model of Sequence Evolution to Evaluate and Improve Phylogenetic Methods

Committee:

____
David M. Hillis, Supervisor

____
James J. Bull

____
Andrew Ellington

____
Robert K. Jansen

____
Beryl B. Simpson

Using a Complex Model of Sequence Evolution to Evaluate and Improve Phylogenetic Methods

by

Mark Travis Holder, B.S.

Dissertation

Presented to the Faculty of the Graduate School of
the University of Texas at Austin
in Partial Fulfillment
of the Requirements
for the Degree of
Doctor of Philosophy

The University of Texas at Austin
December, 2001

Dedication

This dissertation is dedicated to Jim Bull, who has been my model for what a scientist should be.

Acknowledgements

In addition to all of the members of the labs of David Hillis, Jim Bull, David Cannatella, and Tandy Warnow, I would like to acknowledge several scientists who have helped me by generously providing advice or making their source code available: Paul Lewis, Bill Bruno, Aaron Halpern, David Swofford, John Huelsenbeck, and Ziheng Yang. Finally, I would like to thank my wife, Kristina Kichler Holder and my family for their support and encouragement.

Using a Complex Model of Sequence Evolution to Evaluate and Improve Phylogenetic Methods

Publication No. _________

Mark Travis Holder, Ph.D.
The University of Texas at Austin, 2001

Supervisor: David Hillis

The performance of phylogenetic methods was evaluated by testing their success in recovering the true tree from computer simulated data. Data were generated on a variety of tree shapes under a complex model of sequence evolution based on the work of Aaron Halpern and William Bruno. Parameters of the model were estimated by maximum likelihood techniques applied to a phylogeny of 1,610 sequences of mammalian cytochrome *b* genes. These simulations represent a rigorous test of the robustness of phylogenetic methods because several of the simplifying assumptions made by inference methods are violated.

Maximum likelihood methods assuming the general time reversible model of sequence evolution with rate heterogeneity proved to be quite robust, outperforming all other methods on small trees. Distance methods were significantly worse, even when implementing the same model of sequence evolution. On larger trees only distance methods and parsimony techniques were studied. In virtually all cases parsimony outperformed distance-based approaches. The use of simple distance corrections improved performance for four taxon trees and ultrametric sixteen taxon trees but decreased the performance of the neighbor joining method on a 228 taxon tree. Neighbor joining performed as well or better than searches under the minimum evolution criterion in all cases.

In general, the results of these simulations agree with the conclusions of previous studies that phylogenetic methods perform well over a wide range of tree shapes, highly accurate phylogenies for large number of taxa can be obtained from moderate sequence lengths, and model-based distance corrections are much less robust than maximum likelihood implementations of the same model. Maximum likelihood under a common model of sequence evolution was found to be inconsistent on difficult tree shapes when the data were generated under the model described here. Analysis of the spectra of the generating model and the model of inference suggest slight alterations to the general time-reversible model that may improve its performance.

Table of Contents

Chapter 1 - Introduction

The elucidation of the tree of life is a fundamental goal of biology. Phylogenies provide the basic framework to interpret evolutionary history for applications in biogeography, ecology, epidemiology, comparative morphology, and molecular biology. Fortunately in the two last decades, as workers throughout the subdisciplines of biology have come to appreciate the importance of considering phylogenetic history, the molecular biology revolution has made one source of data for constructing trees much easier to obtain. Once sequence data have been collected workers are faced with a bewildering array of analyses that have been proposed to produce trees. This dissertation evaluates phylogenetic methodologies using simulated data from complex models of DNA sequence evolution.

Three general approaches have been used to compare systematic methods: arguing for a method based on first principles, comparing how analyses perform when applied to real data from a tree that is widely agreed upon, and testing methods on simulated data for which the true tree is known because it is specified by the researcher. Examining the theoretical underpinnings of methods has resulted in some of the most significant advances in systematics (e.g., Felsenstein 1978, and Tuffley and Steele, 1997), but such a tact is usually limited to small trees and simple models which are analytically tractable. Using real data from an accepted tree is problematic because if the tree is "known," it is probably an unusually easy tree to infer (but see exceptions to this in Cunningham *et al*., 1994). In the last decade, computer simulation studies have taken a central role in evaluating systematic methods. Increases in computing power have made it feasible to generate and analyze a large number of data sets, and simulations allow the researcher to study performance over a wide range of conditions.

Simulations provide theoretical systematists with a flexible, powerful tool, but their relevance depends upon how well the data generation program mimics the processes acting on real data. The development of models of sequence evolution has been driven by the attempt to capture the effects of the forces of molecular evolution which might complicate systematics. The models have been employed as tools for phylogenetic inference either as the assumed model for maximum likelihood or as the basis for pairwise distance corrections. For these purposes it is important that models have few parameters so that the calculations can be done quickly and the accuracy of the phylogenetic estimate is not hindered by high variance arising from estimating a large number of model parameters in addition to the tree. These simplified models have been co-opted as the basis of simulations. The result is that data generators used for simulation studies produce data that are a simplified version of real data, and that correspond exactly to the assumptions of some of the methods that are being tested. Moving to much more complicated models of sequence evolution provides a more rigorous and realistic test.

The GTR Family of Models

The Jukes Cantor (1969) model (JC) assumes that mutations are fixed in a manner consistent with a Poisson process and that a base changes into any of the other three nucleotides with equal probability. Transitions are known to occur more frequently than transversions, so Kimura's two-parameter model (K2P, 1980) allows the rate of these two classes to differ. Most real sequences significantly deviate from equal base frequencies, and Felsenstein's (F81, 1981) model extended JC to allow the equilibrium frequencies of the nucleotides to vary. Hasegawa, Kishino and Yano (HKY, 1985) and Felsenstein (F84, Kishino and Hasegawa, 1989) combined the K80 and F81 models. The culmination of the single nucleotide models was the general time reversible (GTR) model (Lanave, 1984), which allows for differing base frequencies and different instantaneous rate of substitution for each of the pairs of bases. The only constraint of the model is that of time reversibility (that $\pi_i r_{ij}=\pi_j r_{ji}$ where π_i is the frequency of base i and r_{ij} is the rate of the instantaneous rate parameter of the i to j substitution). Rates of evolution appear to vary dramatically from site to site, and this variation can be added to the GTR-based sequence models by allowing for a certain proportion of sites to be constant, allowing sites to vary continuously in their rate, with the distribution of rates across sites described by a continuous distribution such as the Γ distribution as described by Yang 1994), inferring a different rate for each site in the sequence (Olsen, unpublished), or specifying partitions of the data and inferring a different rate for each partittion.

Weaknesses of Current Models

Although the GTR model with rate heterogeneity addresses the most obvious forces of molecular evolution that could act upon a site, there are several oversimplifications inherent in the model. The three most obvious limiting assumptions of the single nucleotide models are homogeneity of the instantaneous rate matrix across sites (the same underlying model for all sites), stationarity of the model across the tree (a site is subjected to the same forces over every part of the phylogeny), and independence of sites (a change at one site does not affect the evolution at other sites). Each of these assumptions is difficult to avoid in a general, statistically powerful manner.

The assumption that the model of evolution is homogeneous across the sequences could be relaxed by applying different models to different *a priori* partitions of the data. It is not clear that the most obvious partitions (the codon position of a protein-coding gene, or stems and loops for genes that encode RNA molecules with functionally important secondary structures) adequately identify the regions of the molecule that require separate models of sequence evolution. A different model could be applied to every site, but most studies do not have enough data to estimate accurately all of the parameters required by such an approach.

Relaxing the assumption that the model of evolution is constant over the entire tree has been accomplished in several ways. For phylogenetic questions involving deep divergences, the frequencies of the nucleotides often differ significantly between taxa. A general Markov model of sequence evolution, which is not time reversible but allows base frequencies to vary over the tree, has been described (Lockhart *et al.*

1994, and Galtier and Guoy 1998). These methods have not been widely applied because of computational difficulties. LogDet distances (Steele 1994, Lockhart et al. 1994, Lake, 1994) provide a distance correction that is robust to changes in nucleotide frequency over the tree (as with all pairwise distance corrections, it is impossible for the method to automatically correct for unequal rates across sites, but Waddell *et al.* 1995 have shown that removing a portion of invariant sites overcomes this weakness to a large extent). Yang *et al*. (1999) have developed non-stationary models which infer a value of a model parameter for every branch in the tree, but these methods seem to produce pathological parameter estimates indicating that overfitting of the data is a serious problem for these models. Thorne *et al.* (1998), Kishino *et al.* (2001) and Huelsenbeck *et al.* (2000) developed techniques that allow the rate of evolution to change over the tree, either by continuous drift or by discrete changes. A generalization of these techniques to other model parameters may represent an efficient way to allow the model of evolution to change over the tree, but this idea has not been implemented or tested.

Non-independence of sites could result from neighbor effects (mutations affecting more than one site or the nucleotide at one position affecting the mutational spectrum of adjacent sites), important base pairing interactions in the transcribed RNA, or correlations between sites within the same codon. The sheer number of potential interactions between sites makes it infeasible to consider them all in a statistically powerful way. Muse (1995) and Tillier *et al*. (1998) have proposed models to deal with the base pairing interactions of rRNA. Codon models address the non-independence caused by the genetic code (another alternative is to translate protein coding sequences into amino acids and apply amino acid models).

Codon and Amino Acid Models

Muse and Gaut (1994) and Goldman and Yang (1994) independently developed models attempting to capture the fact that synonymous substitutions occur much more frequently than changes which alter the amino acid. Both of these models disallow changes to codons that differ by more than one DNA substitution from the original codon. The Muse and Gaut model adds one new parameter to the GTR family of models. The new parameter is ω, the nonsynonymous to synonymous rate ratio. Muse and Gaut's model can be expanded by adding different equilibrium frequencies for each of the codons. No consideration is given to the similarity of the amino acids in the model, despite the fact that conservative amino acid changes are known to occur more frequently than changes to a residue whose chemical nature is very different from the original amino acid. Goldman and Yang's model can also be used to calculate w, but the model is formulated with the rate of change between two codons determined by the chemical similarity of the amino acids that they code for, as measured by Grantham's distance(1974).

Both the Muse-Gaut and the Goldman-Yang models automatically generate rate heterogeneity and non-independence across nucleotide sites. These models do not incorporate rate heterogeneity across different codons. This constraint can be relaxed by adding Γ-distributed rate heterogeneity, but the other model parameters are still homogeneous across codons.

Amino acid models span the range from very simplistic to extremely parameter rich. Bishop and Friday (1987) introduced a version of the Cavender-Farris model for amino acids (all changes equally likely). Incorporating different equilibrium frequencies for each of the amino acids, the proportional model, is straightforward. Neither of these models capture information about similarity of amino acids, nor do they seem to fit real data well. On the other extreme, every substitution rate can be assumed to have its own parameter. The REV model (Adachi and Hasegawa, 1996) for amino acids has 208 free parameters (as opposed to 8 free parameters for GTR, the nucleotide version of Rev). Constraining the instantaneous rate of change to be zero for all amino acids that do not have codons that are adjacent in the genetic code (referred to as REV0) reduces the number of parameters to 88 (for the vertebrate mitochondrial code). Neither REV nor REV0 have been widely applied in systematics because of the large amount of data needed to infer their parameters.

Yang *et al.* (1998) reviewed the models of amino acid evolution that can be employed for phylogenetic analysis, introduced a useful nomenclature for describing these models, and introduced some extensions to previous models. They distinguish two paths that have been taken to add realism to amino acid models without requiring a large number of nuisance parameters be inferred: empirical and mechanistic approaches. Empirical models are created by comparing a large number of real sequences in a pairwise manner and counting the number of times each type of amino acid substitution occurs then scaling the matrix to a desired divergence. This approach was originally taken by Dayhoff *et al.* (1978) leading to the widely used PAM matrices. Jones *et al.* (1992) have produced an updated version of the matrix (the JTT model) after analyzing a larger data base of protein sequences. Presumably this method would generate reasonable transition probabilities for amino acids, but there are complications. Converting pairwise distances between known proteins into a matrix of transition probabilities requires assumptions about rate heterogeneity between sites and assumes that the same type of changes are occurring at every site (see Wilbur, 1985, for a discussion of the internal inconsistencies in the PAM matrix that may arise from ignoring rate heterogeneity). Adachi and Hasegawa (1996) produced a matrix specifically for mitochondrial proteins by inferring the parameters of the REV model on a data set of the protein-encoding portion of the mitochondrial genome from 24 taxa (mainly mammals). This empirical matrix is called mtREV24, and it (or the very similar mtMammREV matrix) can be used in place of a matrix based on Dayhoff or Jones.

An amino acid model can be built from a codon model by collapsing all of the synonymous codons into one amino acid state. These mechanistic models explicitly deal with the effects of the genetic code on amino acid replacement patterns. If no attempt is made to consider differences in rates depending on the identity of the amino acids, then the model, referred to as the equal-distance model, is essentially an amino acid version of the Muse and Gaut model. Yang *et al.* (1998) analyzed the protein-encoding genes of 20 mammalian mitochondrial genomes. They found that all amino acid models were significantly improved by adding Γ-distributed rate heterogeneity between sites. The mechanistic models gave higher likelihoods than the empirical models, but REV resulted in much higher likelihood than either class of models indicating that neither approach was capturing enough complexity for large analyses. For the empirical models, mtMammREV outperformed JTT and Dayhoff. For the mechanistic models they considered two weighting schemes based on different metrics of amino acid dissimilarity, Grantham's distance (1974) and the distance of Miyata *et al*. (1979). The best performance was obtained from calculating the acceptance rate as an exponential function of Miyata *et al.*'s distance. Unfortunately implementation of the mechanistic models are quite slow, making these methods difficult to study in large simulation experiments.

Site Specific Residue Frequency Model

Clearly different sites in a protein perform different functions. For instance, a residue close to the active site of an enzyme or involved in an important folding interaction might be constrained to be one of the 20 amino acids. Other sites in the protein might be free to be any amino acid. In the middle ground between these two extremes, we would expect some sites in which natural selection (Darwin, 1859) requires that the amino acid be a certain size or have a certain polarity, or "prefers" amino acids based on a mixture of steric and chemical properties. In light of this intuition about constraints on protein evolution, the homogeneous models of amino acid and codon replacement seem greatly oversimplified.

Homogeneous models of amino acid replacement can be given rate heterogeneity, but the notion that slowly evolving sites obey the same rules as the sites with the highest rates of change seems unrealistic. There are two obvious biological reasons for sites to have low rates of change: low mutation rate and high level of constraint. Continuous distributions of rates across sites, like the gamma, treat slow sites as if they simply have a low mutation rate. Variation in mutation rate is a plausible explanation for some heterogeneity, particularly if rate variation occurs between widely separated genes or genes from different genomes. The fact that pseudogenes show considerably less rate heterogeneity than is seen in coding sequences (Yang 1994) implies that differing levels of selective constraint produce much of the rate variation observed in coding sequences. Non-homogeneity has been addressed to some extent by secondary structure specific models of protein evolution (Goldman et al. 1996, Thorne et al. 1996, and Goldman et al. 1998), but not in a way that recognizes differences in constraints between adjacent sites.

Halpern and Bruno (1998) have proposed a model of codon evolution (HB hereafter) that allows each site in the protein to have its own set of amino acid preferences . The model is based on the biologically intuitive approach of treating substitution rates as a function of both a mutational processes and selection. Mutations are proposed according a single nucleotide model (in their paper they used HKY). All of the sites in the protein use the same model parameters to describe the occurrence of mutations. Whether or not a mutation becomes fixed is determined by the fitness of the resulting codon compared to the fitness of the original codon. By assuming time-reversibility and weak selection (s<<1), Halpern and Bruno developed a way to infer the fixation probability of a mutation from codon A to codon B from the equilibrium frequencies of the codons. Assuming that codon bias is the result of uneven nucleotide usage (not selection), the equilibrium frequency of a codon can be calculated from the frequency of the amino acid that it codes for and the nucleotide frequencies. Halpern and Bruno fit their model to 74 viral sequences using rough estimates of amino acid frequencies. Their procedure for estimating amino acid frequencies was to count the number of times each amino acid arises on a phylogeny, add a small number of pseudocounts to the total for each amino acid (whether that amino acid is observed at that site in the sequence or not), and then normalize so that the frequencies sum to one. The purpose of the pseudocounts is to avoid overfitting the data. They showed through simulations that single nucleotide based distance corrections severely underestimate long distances, but a distance correction based on the HB model performed well.

I have re-implemented the HB model in a program that, if given a tree, can infer maximum likelihood estimates of all of the parameters of the model. Because of the large number of parameters that must be estimated the model must be fit to an enormous data set to give reliable parameter estimates (this is why Halpern and Bruno opted to approximate the frequencies and use pseudocounts). I have fit the model to a tree of 1610 unique mammalian cytochrome *b* sequences. Treating the parameter estimates as a reasonable first approximation, the model can be used to simulate data which are much more complex and realistic than data generated in previous simulation studies. Furthermore the data from HB model simulations can be analyzed at the nucleotide, or amino acid level. This presents an opportunity to contrast these levels of analysis in a single simulation study.

Chapter 2 – Model Inference

Overview

The HB model provides a framework for generating simulated data that is much more complex than the models routinely used to analyze phylogenetic data, and a basis for comparing amino acid analyses to DNA based analyses. Simply choosing an appealing model of sequence evolution does not make a simulation study worthwhile. The model must be distinct enough from previously used models so that new insight can be gained, and it must be biologically plausible for it to be relevant to systematists. The HB model meets both of these criteria. For a parameter-rich model, such as this one, an exhaustive sweep over all of the possible parameter space is not feasible. The parameters must be fit to real data in such a way that the model represents a reasonable approximation of the true evolutionary process. The challenges of fitting the model to real data involve finding an appropriate data set and developing software to find the maximum likelihood estimate of the parameters in a reasonable amount of time.

An appropriate data set is one that is likely to match the assumptions of HB model and has enough data to allow for inference of the parameters. The HB model is far too complex for an analytical solution to be possible, so the inference of the parameters must be done heuristically, using numerical optimization techniques. The model presents several computational problems that are not issues for simpler models of sequence evolution. In this chapter I will discuss and justify the criteria I used for choosing a real data set, and then describe the implementation of the model into software.

Description of Parameters

The HB model requires the specification of parameters describing mutation and selection. I implemented the model with a GTR mutational model. This requires 8 parameters (three base frequencies, and parameters for the rate of AC, AG, AT, CG, and CT mutations relative to the GT mutations). As with other maximum likelihood models, branch lengths must be estimated. For an unrooted tree with *n* taxa there are 2*n*-3 branches. Finally one must infer 19 amino-acid equilibrium-frequencies for each triplet in the DNA sequence. If the sequence length (in number of codons) is *k*, the total number of parameters, *p*, that must be estimated is :

$$p= 5 + 2n + 19k$$

Clearly the HB model is much more complex than the GTR family of models (for which the number of parameters is simply 5 + 2*n*). By considering the number of parameters in relation to the total amount of data (*n* times *k*) it is evident that the addition of taxa to the problem provides a better ratio of data to parameters than the addition of more sites. It was vital to find a data set containing a large number of taxa.

The difficulty of an estimation problem is clearly affected by the number of parameters that must be inferred, but how these parameters interact with each other is also very important. For example, when using the GTR family of models, it is often quite difficult to obtain robust estimates of the proportion of invariant sites in a sequence and the shape parameter which describes the distribution of rates among variable sites. These parameters interact with each other strongly because it is possible to explain the same data characteristics (excess of constant sites) by modeling the sites as invariant or modeling extreme rate heterogeneity so that some sites are evolving very slowly. The fact that the branch lengths must also be estimated aggravates the problem. For many datasets there is a wide range of acceptable values for these parameters, because if you constrain the proportion of invariant sites to be higher than the maximum likelihood estimate of its value, increasing the values of the shape parameter will largely compensate and a similar score will be obtained (see Figure 1.1).

Many of the parameters of the HB model do not strongly interact with each other. The mutational model affects all of the sites in the molecule, so the description of the mutational process is potentially conflated with selection. Fortunately, whenever the state of a character is a fourfold-degenerate codon the third base position is evolving neutrally, so the mutational parameters alone control its behavior. This does not mean that the amino acid frequency parameters do not affect the mutational parameters at all, but there is a large portion of the data that allows the effects of mutation and selection to be separated from one another (selection acting at levels other than the amino acid level is ignored in the present implementation). The amino acid frequencies at one site influence the frequencies at other sites only through indirect effects (via causing changes in the mutational and branch length parameters). Because each site has a relatively minor effect on the estimates of the branch lengths and mutational parameters, the problem of fitting amino acid frequency parameters to data is almost like *k* separate 19 parameter optimization problems. In other words, the limited scope of the effects of most of the parameters makes the model easier to fit to real data and results in less variance than a model with an equivalent number of parameters but with more interaction between parameters.

Most model fitting is done during the process of phylogenetic inference, and the phylogeny that is used can have an effect on the model parameters (though the effect seems to be modest for parameters not directly related to the rate of evolution, see Yang *et al*. 1995). As mentioned above, adding taxa to the problem is the most promising way of getting enough data to estimate robustly the parameters of the HB model. Unfortunately the problem of finding a phylogeny gets much more difficult as the number of taxa increases, because the number of possible trees grows quickly with the number of taxa. Given the great computational demands of the HB model it is not feasible to conduct simultaneous model fitting and phylogenetic inference in a maximum likelihood framework. For this study the phylogeny must be treated as if it were known. Thus it was important to find a data set for which there was broad agreement on a large portion of the phylogeny.

The last criterion was to find a gene for which the process of evolution is likely to be stationary because the HB model does not allow variation in the parameters across the tree. If the data set had radically different base frequencies or the evolution of the gene was characterized by dramatic changes in the constraints on different residues (or changes in function), the fit of the parameters to the data might be dominated by artifacts of poor model fit.

Cytochrome *b* Sequences

The mammalian cytochrome *b* gene (henceforth, simply called "cytochrome *b*") fits the criteria of being a conserved gene that is well sampled, from a group that is fairly well understood, and for which there is not strong heterogeneity in base frequencies. I downloaded all of the mammalian cytochrome *b* sequences in GenBank that were at least 1000 bases in length (the full length of the gene is around 1140 bases in most mammals). After removing sequences that were identical to another sequence in the data set and removing all sequences with frameshifts or premature stop codons (in an attempt to exclude pseudogenes incorrectly identified as mitochondrial cytochrome *b*), the data consisted of 1610 sequences. Some species were sampled multiple times, but the identical sequences were culled out, so there was some information about the evolution of the gene in every sequence. The taxonomic coverage of mammals in the cytochrome *b* data set is extraordinarily good for many groups. Chiroptera is conspicuously under-sampled (with representatives of only one genus included), and a few of the less species-rich orders of mammals (Pholidota, Dermoptera, Afrosoricdae, and Scandentia) are not represented at all. Rodents, cetartiodactyls, primates, carnivores, and marsupials are all well represented in terms of total number and inclusion of their major constituent groups. Birds are also well sampled for cytochrome *b*, but they were not examined in this study (in fact, Hastad and Bjorklund used avian cytochrome *b* genes as the basis of a heterogeneous model to study the performance of parsimony and distance methods on relatively small trees). Birds were not considered in this study because the total number of avian and mammalian cytochrome *b* sequences was so large that inference of parameters would have been extremely slow. Rather than select some sequences from each group, I elected to concentrate on the mammalian sequences in the hopes that assumptions of the HB model (namely stationarity) would be less likely to be violated when sequences from only one of the groups were used.

Cytochrome *b* is a mitochondrial gene that codes for part of the electron transport chain. The protein is constrained by several interactions with other proteins in the cytochrome *bc*$_1$ complex and the need to bind to two heme groups. The crystal structure of the complex, with cytochrome *b* at its center, has been determined to high resolution through X-ray crystallography (Zhang *et al.*, 1998). The incorporation of structural information in the form of separate models for different secondary structural regions, or the *a priori* identification of amino acid residues that interact with each other, is an exciting prospect, but beyond the scope of this study.

Phylogenetic Inference of Mammalian Cytochrome *b*

To infer the fit of the HB model a tree is required. The phylogeny of mammals has been the focus of a great deal of research, so the cytochrome *b* data did not have to provide all of the information used in constructing the tree. The strategy that I employed was to provide constraints on relationships between taxa, and then perform parsimony searches on the cytochrome *b* data set. The enormous size of the data set made thorough searches unwieldy. Instead I performed a series of fast bootstrap searches. Groups that had strong bootstrap support in one search were constrained in the next search.

Initially the searches involved only the stepwise addition of taxa, with no branch swapping. As more of the tree became constrained more thorough searching was feasible, and I moved to subtree pruning and regrafting (SPR) swapping, and then to tree bisection reconnection (TBR) swapping. Performing simple searches are unlikely to produce spurious support for nodes in the tree unless there is a strong bias in the tree produced by the stepwise addition process. Addition sequence order was randomized in the searches to limit the effect of this bias.

Because cytochrome *b* is notoriously unreliable for deep relationships, it was important that those relationships be constrained. Delimiting the major groups of mammals has been contentious in the past but there is broad support for most of the major orders (Rodentia, Lagomorpha, Primates, Carnivora, Perissodactyla, Tubulidentata, Proboscidea, Sirenia, Hyracoidea). Many of the recent disputes about higher level mammalian relationships are not relevant to this dissertation because the members of the groups have not been sequenced for cytochrome *b*. For example, the order Insectivora has undergone dramatic changes, namely being split into Dermoptera, Macroscelidea, Scandentia, Eulipotyphla, and Afrosoricidae. Only Eulipotyphlan and one Macroscelid insectivores are sampled for cytochrome *b*.

Molecular studies have made two apparently robust changes to the groupings that have been traditionally referred to as orders: the inclusion of cetaceans within Artiodactyla and the split of Insectivora (after the groups Dermoptera, Macroscelidea and Scandentia had been removed) into Afrosoricidae and Eulipotyphla. Because of taxon sampling in cytochrome *b*, the latter of these changes was not relevant to this study. I constrained Cetaceans to be sister to or within Artiodactyla during the tree searches. The most important contribution of molecular evidence to higher level relationships in mammals has been in the recognition of the groupings of the orders. There is now strong support for Paenungulata (Proboscidea, Hyracoidea, and Sirenia), Afrotheria (Afrosoricidae, Tubulidentata, Macroscelidea, and Paenungulata), Laurasiatheria (Eulipotyphla, Chiroptera, Cetartiodactyla, Perissodactyla, Carnivora, and Pholidota), Glires (the grouping of Rodentia and Lagomorpha, which had also been proposed based on morphology), Euarchonta (Primates, Dermoptera, and Scandentia), Euarchonta-Glires and Euarchonta/Glires/Laurasiatheria (Madsen *et al*. 2001; Murphy *et al*. 2001). For the orders represented in this study, these results almost provide a completely bifurcating tree. The relationships within Laurasiatheria and Paenungulata as well as the rooting of the mammal tree are still unresolved. The tree which provided the parameter estimates that were used in the simulation studies was inferred before the work of Madsen *et al.* and Murphy *et al.*, so the Euarchonta/Glires and Euarchonta/Glires/Laurasiatheria clades (which were uncertain hypotheses until recently) were not constrained in the analyses described below.

All of the tree searches were performed using unordered parsimony and using a stepmatrix that assigned a weight of six to transversions and one to transitions. Initially I conducted exploratory searches to verify that there was not strong signal in the cytochrome *b* data that conflicted with the current understanding of mammalian relationships. These were performed with no constraints at all or with a “deep” level constraint shown in Figure 2.1. Note that this constraint has no structure among the orders of eutherians, and Eulipotyphla was not constrained to be monophyletic. As expected the relationships inferred in the unconstrained analyses were not congruent with the monophyly of several of the recognized orders, but none of these results showed strong bootstrap support.   After ten searches without strong bootstrap support for any set of unexpected relationships, a new backbone constraint of early mammalian relationships (shown in Figure 2.2) was used. This constraint used more information from recent studies of mammalian systematics (as summarized by Waddell *et al*. 1999) to resolve the deep parts of the tree.

Starting after the second round of searches, the constraint tree for each search contained the higher level relationships shown in Figure 2.1 or Figure 2.2 as well as nodes with high bootstrap support in the previous round. For a branch’s support to be considered high enough to constrain in the next analyses, the bootstrap proportions had to exceed a cutoff value in both the unordered and the transition/transversion searches. Table 2.1 shows the history of the searches, including the basic constraint tree used, the cutoff for branches in the previous round to be considered well supported enough to constrain, and the type of swapping employed. The type of swapping was not always identical between the unordered and the transition/transversion analyses, because searches with step matrices are slower and I was attempting to keep the running times low. It is important to realize that the bootstrap values in these searches may be significantly inflated in this strategy. A node with an asymptotic bootstrap proportion of 85% might easily have an observed bootstrap proportion of 90% in one run, and the chances of this are obviously increased by running multiple analyses. As soon as a node exceeded the cutoff level it was constrained in subsequent searches. As more and more of the tree became constrained, bootstrap proportions were likely to be inflated because some possible topologies were prohibited. Given that I needed only a point estimate of the topology, not measures of support, this strategy seemed to be a valuable way to include all of the data in the analysis, but not spend too much time on unrealistic trees or rediscovering the same well-supported clades.

After the number of new nodes being constrained in each round had decreased and the cutoff level had been lowered, 200 TBR searches were performed under unordered and transition/transversion weighting. The most-parsimonious tree (under transition/transversion weighting) was chosen as the tree for model inference. After the topology was chosen, a misaligned region of the marsupial data was found. There are very few positions of ambiguous alignment in cytochrome *b*, because there is almost no length variation except in the last few amino acids of the protein, and the misaligned region only affected a few bases for a few taxa. To verify that this alignment problem did not strongly affect the topology, a pair of bootstrap searches were performed with no constraints on the Marsupials. The resulting tree, which was used for the inference of parameters, had no differences from the previous analyses for the clades with bootstrap support of over 50%.

Description of the Tree Used For Inference

Although unconstrained analyses of the cytochrome *b* sequences failed to reconstruct the oldest divergences in mammals which are considered strongly supported (on the basis of other data), the molecule did seem to recover many of more recent groups with strong bootstrap support from even the earliest analyses. The tree used for inference had a few unconventional (and probably unreliable) groupings, but they typically involved deep divergences in the tree; so even if these groupings are erroneous it is unlikely that they had a substantial effect on parameter estimates. The discussion below highlights the most questionable groupings, but most of the phylogeny was in strong agreement with the current understanding of mammalian relationships. It is possible that a strategy of more rigorous constraints would have produced a more reliable tree, but given that there are few uncontested relationships such constraints are not guaranteed to be correct. Figure 2.3 shows the relationships between the major groups in the tree that was used for inference.

Most of the well supported groups of Marsupials relevant to this study (Didelphidae, Caenolestidae, Peremelemorphia, and Dasyuromorphia) were recovered. The enigmatic *Dromiciops gliroides* grouped inside Diprotondotia (which had poor taxon sampling and therefore long branches). Didelphids were sister to the rest of Metatheria, despite Colgan's (1999) conclusion that the Ameridelphia (Didelphids and Caenolestids) are sister to all other marsupials. The placement of *Dromiciops* and Didelphidae are the most troubling aspect of the Marsupial clade. Within Dasyuromorphia, *Myrmecobius* was sister to the Sminthopsinae. *Planigale* fell within *Sminthopsis*, as found by Painter *et al*. (1995), but *contra* Blacket *et al.* (1999). Relationships within Phascogalinae differed from Armstrong (1998) only by a change in the rooting of *Antechinus*. Dasyurinae was paraphyletic with respect to the Phascolosoricinae.

Within rodents the following major groupings were found: Sciuridae, Myoxidae, Geomyidae, Heteromyidae (and the well supported grouping of Geomyidae and Heteromyidae), Hystricognathous families, and most of Muridae.

Muridae was monophyletic with the exception of *Tachyoryctes splendens*, a fossorial rodent from Africa that has been placed (along with other African and Asian species that were not in this data set) into the family Rhizomyidae. *Tachyoryctes* was sister to *Zapus trinotatus*, the only Dipodid in the data set. This grouping is questionable; for this study *Zapus* should probably be sister to Muridae with *Tachyoryctes* as the basal Murid (R. Adkins, personal communication). Within murids the expected groups - the Arvicolinae, Cricetinae, Gerbellinae, Otomyinae, Cricetomyinae, the South American sigmodontines, the “neotomyine” sigmodontines (Engel *et al*., 1998), and a clade of “acomyines” (Dubois *et al*., 1999) – were recovered. The subfamily Nesomyinae appears to be a polyphyletic assemblage of two clades of murids, and Murinae was paraphyletic. As expected, Cricetinae was sister to Arvicolinae, and the Sigmondontines were sister to this clade, although the rooting differed from Dubois *et al*. (1999).

Among the Hystricognaths, the relationships were in accord with Nedbal *et al.* (1994), Catzeflis *et al. (*1995), Lara *et al. (*1996), and Lessa *et al.* (1998). The present study include porcupines and *Myoprocta*, which were not included in those studies. Pruning these taxa from the tree results in Bathyergidae sister to the other hystricognaths, with *Cavia* sister to the Octodondoidea. Within Octodondoidea, Echymidae was sister to the a clade of Ctenomyidae and Octodontidae.

The relationships between the major Cetartiodactyl groups was unconventional (and in light of strong molecular evidence probably wrong). The major groups Camelidae, Suiformes, Cetacea, Hippopotomidae, and Ruminatia were found, but the well-supported relationships between these groups were not recovered Camelidae was the basal group as expected, but instead of the group “Whippomorpha” (Cetacea and Hippopotomidae) being sister to Ruminatia, they were inferred to be sister to Suiformes and Ruminantia (although when considering the branch lengths, the branching arrangement was close to a polytomy). Within these groupings the tree was quite congruent with other estimate of phylogeny.

Tragulus was the basal lineage of ruminates with Giraffidae, Cervidae, Moschidae, and Bovidae all recovered. All of the well supported groups in Bovidae found by Matthee and Davis's (2001) analysis based on four nuclear genes and three mitochondrial genes were found. As with other molecular phylogenies, the Odontoceti were inferred to be paraphyletic (*contra* morphological evidence), but the Mysteceti were monophyletic. Relationships were within Delphinids were very similar to those found by LeDuc *et al*. (1998, using the same cytochrome *b* seqences).

The major groups within Carnivora were found. Herpestidae was sister to Felidae and this clade was sister to the Arctoidea. Canidae, Mustelidae, Mephitids, Pinnipeds, and Ursidae were all monophyletic. Questionable results include the failure to reconstruct a monophyletic Procyonidae and the placement of the enigmatic *Ailurus fulgens* as sister to Ursids (see Flynn *et al.* 2000).

Finding Maximum Likelihood Estimates of the Model Parameters

Even assuming that the tree is known, the fitting of the Halpern Bruno model to the cytochrome *b* data is an enormous computational task. The problem amounts to a simultaneous maximization of 10,369 parameters given approximately 1,800,000 nucleotides. Clearly the maximization is beyond the scope of analytical techniques, so numerical optimization following Powell (1964) was employed. In this technique, a point is scored, and then the optimal value of each parameter is determined by maximizing the function along a line. Initially, the direction of each line is simply an change in one parameter with no change in the other independent parameters. Multiparameter directions can be added to speed the approach to the optimum.

Fortunately approximations are available for all of the parameters, so that the starting point for the inference is far from random. To obtain starting values of the mutational parameters the tree was scored in PAUP* (Swofford) under the HKY model with gamma-distributed rate heterogeneity. After one week of computation, the inference seemed to be making only slight changes to branch lengths (based on the parameter iteration log and the rate of change of the likelihood), and the run was terminated. This analysis also provided starting points for the branch lengths. For the 7,144 amino acid frequency parameters, an initial approximation was developed in which the amino acid's frequency was equal to the proportion of the tree that was inferred to have that amino acid (ancestral character states were inferred using the parsimony criterion). The proportion was determined by weighting each branch by its branch length (using the HKY+gamma branch lengths).

At most sites in the data set at least one of the twenty amino acids was absent. This greatly simplified the parameter optimization. If an amino acid was never observed and none of its codons are mutationally between two amino acids with non-zero frequencies, then the amino acid was assumed to have a maximum likelihood estimate of zero. This simplification was a very important component of the optimization because it greatly reduced run time in two ways. First, amino acids with assumed zero frequencies represent parameters that do not need to be optimized. Second, the number of states for a site without a particular amino acid is reduced by the number of triplets that code for that amino acid. The computational time of maximum likelihood inference on a tree generally scales as the square of the number of states (at each node the calculation involves multiplying the probability of going from each of the *n* states in the ancestor to each of the *n* states in the descendant times the conditional likelihood of the tree below the descendant). The time needed to calculate the *n* by *n* matrix of probabilities is usually negligible compared to the *n*-squared multiplications involving the matrix. For the HB model each site has a different model, so for each branch on the tree the probability matrix must be calculated for each amino acid. The calculation of the probability matrix from the eigenvalues and eigenvectors of the model and the branch lengths scales as the cube of the number of states. The inference program spent approximately 95% of its time in the function which calculates the probability matrix, so this step was clearly the rate limiting step of the inference.

Because the HB model is a codon model (with the stop codons disallowed) the number of states can be as high as 60. In the extreme case of a site that is fixed for an amino acid with only two codons, the maximum likelihood estimate of the frequency of that amino acid is guaranteed to be one. By assuming this from the start, 20 parameter optimizations are avoided and every time the site is scored (as it must be when mutational parameters or branch lengths are changed) the computation takes approximately 1/27,000 the time (because the number of states is reduced 30 fold and the rate limiting step scales with the cube of the number of states).

Branch length optimization is traditionally achieved by sweeping over the tree and optimizing each branch in isolation. The Newton-Raphson single parameter optimization strategy works well, but requires the calculation of the first and second derivatives of the likelihood with respect to the branch length. This calculation is possible under the HB model, but it involves the calculation of three matrices. The run time of each of these calculations scales as the cube of the number of states. Because the HB model has a much larger number of states than the “standard” nucleotide models, even after removal of amino acids with an empirical frequency of zero, this step is likely to be costly. Instead optimization of each branch followed Brent’s (1973) method of linear optimization, which tries to find the optimum by repeatedly fitting a parabola to the known values and moving to the apex of the parabola. This method only requires the calculation of the probability of change matrix.

After four rounds of parameter optimization, branch length optimization was performed. In subsequent rounds the decision of whether to optimize branch lengths or model parameters was made by repeating whichever technique provided the larger increase in likelihood score the last time it was performed. Figures 2.4 and 2.5 show the increase in likelihood and decrease in the magnitude of change in the parameters over the 38 rounds of optimization.   Optimization was stopped when the change in Ln likelihood was less than 0.05 following a full round of parameter optimization. Undoubtedly the likelihood would have continued to improve, but the shape of the optimization curve (Figures 2.4 and 2.5), indicates that further computation time is probably unwarranted given the rate at which the likelihood and parameter values are changing in the last stages of the optimization.

Program Requirements

The entire parameter optimization procedure took approximately two months of computational time on a 677 MHz G4 processor with 512 megabytes of physical RAM. Much of the challenge of writing the software involved minimizing the memory requirements. Use of virtual memory or other techniques of storing information on the hard disks as opposed to in physical memory, dramatically increases run times. To avoid using such techniques many data structures shared workspaces for calculations. Despite these efforts the program still required approximately 440 MB.

Likelihoods for individual sites on a tree of 1610 taxa can easily become too small to store in primitive C/C++ datatypes. This problem, known as underflow, was avoided by multiplying the conditional likelihood of each tip by a constant factor (a separate multiplier was used for each site, based on preliminary calculations). In some optimization rounds, during the pass down the tree which calculates the likelihood, checks were made to assure that at least one of the conditional likelihoods for a character state was 50 decimal places above the cutoff for the loss of precision due to underflow. The scaling procedure seemed to work well for this inference, but may not be a robust way of dealing with underflow in software in which many trees must be evaluated.

Chapter 3 – Simulations – Four-taxon Tree

Overview of Simulations

The type of tree on which data are simulated can radically alter the conclusions from a simulation study. An obvious example of this phenomenon is the case of long-branch attraction. Felsenstein (1978) described a tree that will positively mislead parsimony (Figure 3.1a). Stochastic homoplasy leads to convergence between the two taxa that are at the ends of long branches, known as long-branch attraction. Simulations of the trees of this general shape typically show likelihood and distance-based criteria dramatically outperforming parsimony, because those methods are able to recognize the presence of long branches, and account for their effects. A slightly different tree (Figure 3.1b) has the two taxa with long branches being sister to each other. Simulations based on extreme versions of this tree will reach the opposite conclusion. In these cases parsimony's interpretation of every shared character as a phylogenetic signal helps it, because the noise in the evolutionary process agrees with the true phylogenetic signal. Parsimony will reconstruct the true tree with less data than is required by maximum likelihood methods.

Clearly, it is necessary to interpret the results of a simulation in the context of the experiment and try to steer clear of conclusions such as "parsimony is better than likelihood" (or the opposite), on the basis of simulations from one type of tree. Many phylogenetic methods have biases in favor of one topology compared to another; if the model tree in a simulation agrees with a method's bias, it is easy to interpret the result as the method performing well (Bruno and Halpern, 1999).

To avoid spurious conclusions that arise from only examining one type of tree, I have performed four simulation experiments using the HB model, fit to cytochrome *b* sequences, as a data generator. The first experiment is the most thorough in terms of methods and models examined. It is a replicate of the studies of Huelsenbeck and Hillis (1993), Gaut and Lewis (1995), and Huelsenbeck (1995a, 1995b) on four-taxon trees. A study of a group of contrived sixteen-taxon trees was designed to test the discriminatory power of likelihood, parsimony, and distance-based approaches.   The third experiment is based on a 228-taxon tree examined by Hillis (1996) to study large-scale phylogenetics.   Finally, trees generated under a random branching process, the Yule process, were examined as a way of looking at large trees in a wide range of tree shapes.

Four-taxon Trees

Unrooted four-taxon trees and rooted three taxon trees, have played a prominent role in the history of systematic theory. Exhaustive tree searches can be performed (so the results are not clouded by uncertainty as to how well the particular tree searching routine is solving the problem of finding a tree that satisfies the optimality criterion). Every biological tree that is inferred can be reduced to a series of four-taxon statements, in which the leaves of the tree are no longer always observed taxa, but might instead be reconstructed ancestors. Quartet methods take this rationale to the extreme by breaking every problem into four taxa sets, computing a phylogeny for each of these quartets, and then assembling a full tree. Thus, four-taxon trees could be considered the fundamental problem in phylogenetics.

Huelsenbeck and Hillis (1994) simplified the parameter space of the four-taxon tree problem by constraining two non-sister terminal branches to be one length and the other three branches on the tree to be a separate length. This allows a large range of trees to be examined by covering just two parameters (Figure 3.2). The lower left corner of the parameter space is the easiest tree shape to correctly infer, with all of the branches being short and the total amount of noise being low. As one moves to the right on the graph the amount of phylogenetic signal (length of the internal branch) increases, but noise also accumulates. Moving up the graph, noise spuriously uniting two taxa increases, with no increase in the amount of signal. The upper left corner is the most difficult, representing the tree that Felsenstein described for which parsimony is positively misleading. Huelsenbeck and Hillis studied the performance of a variety of methods over this parameter space when the model of evolution used to generate the data was quite simple. Their general conclusions were that most methods performed quite well, with the exception of UPGMA and Lake's method of invariants. Gaut and Lewis, repeated these simulations, but inferred trees using maximum likelihood. They concluded that maximum likelihood also performed well over most of the parameter space, and seemed fairly robust to violations of the assumptions of the model of evolution. Huelsenbeck extended his study with Hillis by investigating maximum likelihood and a wider range of distance-based approaches. He found that maximum likelihood techniques outperformed all other methods investigated, but the differences between maximum likelihood and distance-based approaches was not huge when distances were corrected using model-based methods.

I have repeated these simulations using the Halpern/Bruno model fit to cytochrome *b* data. The only difference in the simulations other than the model of evolution, was the way that branch lengths were specified. Previous workers expressed branch lengths in the expected percent divergence between the sequence on one end of the branch and the sequence on the other end. This has the advantage that when methods are examined from 0% divergence up to 75% divergence, the entire range of possible parameter values has been covered. There are two disadvantages to this parameterization scheme. First, this way of expressing branch lengths is not commonly used in any other part of phylogenetics. In most contexts, model-based branch-length estimates are expressed in the expected number of changes per site. Other than just being more common, this description is easier to apply to a wide range of models (when base frequencies are not equal the maximum percent sequence divergence is not 75%, and so each model has its own maximum value making branch lengths in expected percent sequence divergence hard to compare between models). Second, it is easier to extrapolate from branch lengths expressed in expected changes per site. This number should be roughly correlated with the branch length when expressed in time; so if one were interested in the performance of a method when the divergences were twice as old, doubling the branch length should provide the answer. This is valid if branches are expressed in changes per site, but not if they are in percent sequence divergence. A side effect of this, is that though the previous studies cover branch length space exhaustively, the visual impression is that long branches (those over 50 % divergence, for example) occupy a fairly small part of parameter space. However if the results are plotted as a function of the expected number of changes per site with an arbitrary (but large) maximum such as 5.0, then the majority of parameter space is made up of long branches. Clearly systematists would try to sample taxa such that enormous branch lengths are avoided, but when presenting simulations, it seems preferable to choose axes that do not magnify or minimize regions of parameter space. I varied branch lengths from 0.05 expected changes per site up to 1.0 expected changes per site in twenty steps. With two branch lengths being varied, there were a total of 400 simulation conditions. One hundred replicate data sets were produced for each condition for a total of 40,000 simulations.

DNA Analyses - Parsimony

Because scoring of four-taxon trees is relatively fast, a large number of methods could be examined on these simulated data sets. I examined DNA parsimony using three different weightings schemes: unordered parsimony, weighting transversions 1.8 times more than transitions (based on an estimate of the instantaneous transition to transversion rate ratio), and using an asymmetric stepmatrix based on the instantaneous rate matrix inferred from one simulated data set using the GTR model of sequence evolution (this analysis will simply be referred to as weighted parsimony).

DNA Analyses - Maximum Likelihood

Nucleotide-based maximum likelihood techniques were investigated by implementing GTR with rate heterogeneity. A full process of model selection was not done, because of time constraints. Subsequent comparison of the HKY model of sequence evolution to the GTR model indicate that only 58 of the 40,000 replicates would have preferred HKY to GTR (based on a likelihood-ratio test assuming that the test statistic follows a chi-squared distribution), thus it is unlikely that insufficient attention to model selection seriously affected the results (although there are a number of models of complexity intermediate between the GTR model and HKY, and these models were not tested). For analyses of the simulated data sets, I used likelihood-ratio tests to select the preferred model of among-site rate variation. If both types of rate heterogeneity (the GTR+Γ+I model) provided a statistically better fit than either type of rate heterogeneity by itself then both were used, otherwise whichever one parameter model provided a higher likelihood was used. Using this methodology 52 replicates preferred both types of rate heterogeneity, 14,973 preferred the gamma rate correction only, and the rest (24,975 replicates) preferred the assumption of a class of invariant sites. These likelihood analyses are referred to as the preferred model analyses. Analyses using only gamma rates and using only invariant sites for all replicates were also collected in the process of model testing. Surprisingly the analyses simply assuming invariant sites performed slightly better than using the likelihood-ratio test to select which model was preferable.

Rate heterogeneity can also be dealt with by dividing the data into *a priori* classes of characters that are assumed to have different rates. For coding sequences, such as those being simulated in this study, an obvious choice of partitions was the three codon positions; the name GTR plus site-specific rates will be used to refer to analyses in which a rate of evolution was inferred for each of the codon positions (all sites within that category were assumed to follow that rate).

DNA Analyses - Distance Methods

Distance-based approaches were studied when the distances were altered by a GTR-based correction for multiple hits. Among-site rate variation is known to affect distance estimates, but when estimating pairwise distances, it is not possible to estimate rate heterogeneity across sites.   Fitch-Margoliash and minimum evolution criteria were examined using a correction based on a gamma distribution of rates, with the shape parameter estimated by maximum likelihood from the most parsimonious tree. To avoid conflating the results of distance approaches with errors introduced from parsimony or by simply assuming gamma rates, the popular minimum evolution criterion (ME hereafter) was also studied using the GTR model with the rate heterogeneity correction that was selected as the preferred model (as described above). For this model the rate heterogeneity parameters were estimated from the true tree (as opposed to the most parsimonious tree).

When accounting for rate variation across sites by using the invariant sites correction, some proportion of sites is essentially removed from the data matrix. Which sites are removed can affect the analyses so the preferred model distance corrections were done two ways: removing constant sites of a particular base in proportion to that base's frequency in the whole data matrix, and removing constant sites from each base by that base's frequency among just the constant sites. These two approaches are referred to ME-all and ME-constant respectively. In many simulation studies there would be no basis for expecting a difference between these two approaches, but in the HB simulations the frequencies of the bases at constant sites are not the same as the base frequencies for rapidly evolving sites. The main reason for this effect is that selection maintains some bases because of the amino acid for which they code. Highly constrained amino acids are unlikely to change, and they are also immune to mutational pressure. Third base positions, on the other hand, are rarely constant sites, and mimic the mutational base composition biases much more closely. This property is also true of the real mammalian cytochrome *b* sequences which have much more extreme base composition bias in the third base positions than in the other two.

Amino Acid Analyses - Parsimony

Analyses of amino acid sequences were conducted using parsimony and maximum likelihood techniques. Four amino acid weighting schemes were used for parsimony: unordered, PAM1, PAM250, and a mutational distance matrix. Unordered analyses simply assume that any amino acid changes into any other amino acid at an equal rate. PAM matrices are widely used by molecular biologists in protein alignments. These matrices are based on replacement rates of one amino acid by another as inferred by comparing many pairs of homologous proteins (Dayhoff, 1978). The PAM1 matrix can be viewed as the most appropriate matrix for very similar proteins. Different members of the PAM family useful for more divergent protein sequences can be created by treating the amino acid replacement as a Markov process, and raising the matrix probabilities of change for the PAM1 matrix to an arbitrary power (the branch length). The PAM matrices are usually used as log–odds matrices of costs of any particular replacement. In this context they differ only slightly from stepmatrices used in phylogenetics: different amino acids have different mutabilities and this implies that there is cost associated with not changing state. To produce a matrix for use in standard phylogenetic analysis, I scaled subtracted the diagonal element of each row from every other element in that row (to remove any penalty for no change, but keep the same relative costs of replacements), and then forced the matrix to obey the triangle inequality. I did this for the PAM1 and PAM250 matrices and used them as stepmatrices in parsimony. Finally, I examined the performance of a stepmatrix designed to assign costs of amino acid replacements based on how many nonsynonymous mutational steps are required to change from one amino acid to another. This type of matrix was originally proposed by Felsenstein, as the ProtPars option in his PHYLIP software package. Synonymous changes are assumed to happen so quickly that they can be ignored; for instance it takes three mutations to change from a codon for histidine (CAY) to a codon for methionine (AUR), but this amino acid replacement receives a cost of only two, because the third position change can be a synonymous change if the mutational pathway goes through leucine (CUN). I modified Felsenstein's matrix so that it applies to the vertebrate mitochondrial genetic code.

Amino Acid Analyses –Maximum Likelihood

Maximum likelihood analyses of amino acid sequences were done using PAML (Yang). Given the short sequence length of the simulations (376 amino acids), complex models, such as REV, are not practical. I examined the performance of the proportional model which requires the estimation of 19 free parameters (the amino acid frequencies). The rate of substitution is assumed to be proportional to the destination amino acid's frequency. The other model I studied was the mitochondrial mammalian reversible model (mtMammREV), which is simply a general reversible model of amino acid replacement, but the rate parameters are assumed instead of being inferred. The values of the parameters come from Adachi and Hasegawa's (1996) analysis of entire mammalian mitochondrial genomes. This matrix is suggested when there are not enough data to infer the values of all of the parameters of the REV model. A slight modification of the mtMammREV model involves the estimation of amino acid frequencies as free parameters, but retains Adachi and Hasegawa's rate multipliers for the substitution types. I scored trees under both incarnations of the mtMammREV model, and report the result of the preferred model (based on likelihood-ratio tests) as the empirical model of amino acid evolution. Only 101 of the 40,000 replicates preferred the more complex model of treating the amino acid frequencies as free parameters. Both the proportional and the empirical models of evolution were implemented using gamma-distributed rate heterogeneity (approximated by eight discrete categories).

Four-taxon Results

Figure 3.3 shows the performance of the 16 methods over the entire range of parameter space. There was substantial variance in performance with maximum likelihood on nucleotides sequences (GTR with invariant sites) preferring the true tree on 91.5% of the 40,000 replicates while one version of minimum evolution (using GTR corrections with the preferred model of rate heterogeneity and proportions of invariant sites to be removed estimated by the base frequencies from all sites) only preferred the true tree 68.9% of the time. Figures 3.4-3.20 show the performance of each method graphically, with black squares indicating replicates in which the true tree was found in 95 or more of the 100 replicates. When the percentage of correct replicates was less than 95% the box is shaded with a color according to the scale above each graph (pure red represents 0% correct). Replicates in which the true tree was found <33% of the time (worse than guessing) are outlined in white. In the following discussion statistical significance is judged by comparing the total number of replicates that were correctly inferred. General conclusions such as evaluations of parsimony vs. maximum likelihood are always based on comparisons of the best version of each of the methods (with the exception of the GTR method, in these cases GTR with the preferred rate heterogeneity model was used instead of using invariant sites because the latter model was not originally considered to be a method to be investigated). Giving a method partial credit for tree topologies with identical scores alters the results very little (mainly boosting the performance of the unordered parsimony methods).

DNA – Parsimony Methods

The results of this study pertaining to parsimony are qualitatively similar to previous work. The use of stepmatrices dramatically increases the number of times the true tree is inferred (Figures 3.5 and 3.6 compared to Figure 3.4). While parsimony methods do remarkably well over most of the parameter space, the performance drops off sharply as trees begin to resemble those described by Felsenstein. The four-taxon simulations under the Halpern-Bruno model once again confirm that under Markov models of sequence evolution (even complex ones), parsimony is quite robust over a wide range of conditions, but also very susceptible to long-branch attraction.

Parsimony makes no attempt to detect long branches or weigh the possibility of change along a branches according to the branch's length. Whether this is a damning fault depends on whether branch lengths "exist" as evolutionarily meaningful parameters. In other words, if the probability of change at one site along a branch is completely uncorrelated to the probability of change at another character along that branch, then parsimony is justifiable as a maximum likelihood estimator of the phylogeny (Tuffley and Steel, 1997). The simulations performed in this dissertation assume that branches have a meaningful length – a parameter that correlates strongly with the probability of change at every of site in the molecule. When branch lengths become important parameters (because they vary dramatically from branch to branch) parsimony can fail spectacularly.

One could argue that even though the HB model is very complex the true evolutionary process is much more complex and really resembles the "no common mechanism" model described by Tuffley and Steel, and the only reason parsimony fails is because the simulations have an unrealistic set of branch lengths that apply to every site. If the no common mechanism model were true, then the branch lengths inferred from one set of characters should be uncorrelated with the branch lengths inferred from a second set of characters. This is an empirical question, but the Tuffley/Steel model seems unlikely to be realistic for molecular sequence data sets in which the amount of evolutionary time is a rough correlate of the probability of change for most characters. Figure 3.21 shows the correlation between branch lengths estimated from the even numbered characters of the real cytochrome *b* sequences in artiodactyl species and the branch lengths estimated from the odd numbered characters (the tree used is believed to be the maximum likelihood estimate of the topology, the model of evolution used during the inference of branch lengths was GTR with gamma-distributed rate heterogeneity and invariant sites with all parameters estimated from the full data set). Clearly the branch length estimates are highly correlated indicating that it is reasonable to perform simulation of molecular sequences assuming there is a shared branch length for all sites.

DNA – Maximum Likelihood Methods

Analyses of nucleotide sequences using the GTR model of sequence evolution with rate heterogeneity outperformed all other approaches. While the comparison of likelihoods to select a model of rate heterogeneity did produce a method (Figure 3.7) that outperformed parsimony and distance methods, simply using invariant sites (Figure 3.9) to deal with variation in rates performed better than any other method. Modeling rate variation by *a priori* categories was also effective, but worse than using gamma rates or invariant sites only (see Figure 3.10). It may seem counterintuitive that the less biologically motivated methods of dealing with rate variation appear more powerful, but apparently there is enough variation in rates among each codon position, that simply treating each category as having a separate rate is too restrictive.

It is very encouraging that the GTR models are performing well in a simulation in which the data are generated under a model that violates virtually every assumption of the model. Clearly the model is quite robust to violation of its assumptions, however there are several important caveats. First the Halpern/Bruno model, while complex is still an oversimplification of evolution. It assumes stationarity, and the lack of selection for different codons may cause the evolution of the third base positions in the model to evolve in a way that is unrealistically similar to the assumptions of GTR. In fact, approximately one third of the third base positions are evolving under GTR in the HB model because at some sites the only amino acids that occur have fourfold degenerate codons so there is never any selection on the third base position.

Figure 3.22 compares the performance of maximum likelihood to parsimony. For much of the parameter space there is no discernible difference between likelihood and parsimony. In accordance with theory and previous simulation studies, likelihood is outperforming parsimony for trees near or inside the “Felsenstein zone.” Unlike previous studies, the superiority of maximum likelihood does not hold for the most extreme cases of long-branch attraction. Despite the fact that maximum likelihood is not outperforming parsimony over the entire parameter space, these simulations provide little reason to support a parsimony tree over a maximum likelihood tree. There are some indications that parsimony may be slightly outperforming likelihood when all of the branches are long, but this effect is clearly very small if it is real. The converse is not true; there are clearly tree shapes for which maximum likelihood, even under an unreasonably simplistic model, significantly outperforms parsimony. Whether these relatively rare regions of parameter space are enough justification for workers to invest the time required for likelihood tree searches depends on one’s view of what are biologically reasonable branch lengths.

Although the GTR models are performing well over most of the parameter space, they are failing dramatically on trees with extreme long-branch attraction problems. In fact, maximum likelihood under the GTR model with gamma-distributed rate heterogeneity and invariant sites is inconsistent for tree topology estimation when data are generated under the HB model on the most extreme tree examined (three branch lengths of 0.05 and two branch lengths of 1.0). Consistency proofs are difficult for the GTR model with rate heterogeneity, so this conclusion comes from an approximation to infinite data. The expected proportions of all of the data patterns can be calculated for any tree. A data set was produced with all 256 of the possible data patterns for four taxa. Character weights proportional to the expected frequency of that data pattern in an infinite sample are applied to each pattern, and maximum likelihood scores of the three tree topologies are estimated. GTR with rate heterogeneity preferred the wrong tree under this approximation to infinite data. This was true for analysis of the whole simulated sequence, or when the data patterns were calculated for each of the base positions separately. This method is an approximation because PAUP* will only accept integer weights for characters in likelihood analyses, so there is some rounding error. By assigning large weights to all of the data patterns, the rounding error for any single data pattern was limited to a maximum of 1.5e-06.

It has long been recognized that the consistency of maximum likelihood requires that the assumptions of the model of evolution be met, so the inconsistency of GTR with rate heterogeneity is not shocking, but it was not a foregone conclusion either. Going into this study it was entirely unclear whether the simple nucleotide models incorporated enough information about the nucleotide substitution process, that they would correctly estimate the tree over the entire range of trees examined in this study. The minimum internal branch length was 0.05, which equates to approximately 55 changes that could potentially provide phylogenetic signal. The long branches on the inconsistent tree were certainly quite long (1.0 expected changes per site), but the topology is not impossible to infer. This is not a case of branches being so long that no method could possibly succeed. To demonstrate this I have inferred the tree topology under the Halpern/Bruno model on a few of the simulated data sets from the extreme Felsenstein zone trees. I supplied all of the parameters to the model, so this method of inference is not a realistic option, but the true tree is recovered in about 70% of the replicates. This probably represents an upper bound on the performance of any method on these simulations, but it is clearly far better than a method could do by chance. I have also simulated data on the same tree but under a GTR model with gamma-distributed rates and invariant sites. The parameter values were selected to mimic the difficulty of the cytochrome *b* sequences as much as possible (same length with model parameters inferred from a large Halpern/Bruno simulated tree). Under these simulations GTR with rate heterogeneity infers the tree correctly in 76 out of 100 replicates (when data were simulated on the Halpern/Bruno model for this tree, the best GTR implementation only recovered the true tree seven times under these conditions). These results indicate that the poor performance of the methods in the extreme Felsenstein zone is not due to branches being so long that all of the synapomorphies are lost.

The source of the maximum likelihood errors on these extreme cases appears to be underestimation of branch lengths. Cursory examination of the data pattern produced by the Halpern/Bruno model indicates that the non-homogeneity of the model can have serious effects on the abundance of data signals that the GTR models are using to detect long branches. Under the Halpern/Bruno model some sites have such extreme biases that they are essentially two state characters (often with one state being much more common than the other). Which two states are present depends on the particular site in the protein. The implications of this are that the sequences saturate much below 75% sequence divergence, and when the true distance is long it is dramatically underestimated by the GTR models. This underestimation of branch lengths makes it difficult for the model to account for all of the homoplasy in the extreme Felsenstein zone simulations, and the models interpret the homoplasy as true signal. Ways to address this problem will be proposed in Chapter 7.

DNA – Distance Approaches

The results of the distance analyses (Figures 3.11-3.14) stand in stark contrasts to the maximum likelihood results, even in cases in which the model of evolution is the same for the two approaches. Figure 3.23 compares the performance of maximum likelihood and distance using GTR with the preferred model of rate heterogeneity (invariant sites are removed according to the base frequencies in the constant sites). For the distance analyses the rate heterogeneity parameters are being estimated from the true tree, which amounts to an unfair advantage for the distance methods (when conducting a real analyses, an investigator would have to use some method of guessing which tree's parameter to use). There is a large portion of parameter space for which distance methods perform fairly well, but in most cases they are worse than maximum likelihood. In seven of the 400 conditions, distance methods recovered the true tree in one more of the 100 replicates than likelihood; in sixteen conditions both methods were identical; and in the other 377 conditions maximum likelihood did better (for one set ML recovered the true tree 58 times more than ME). There seems to be much more variance in the quality of the answers given by the distance approaches, so that even under conditions far from the Felsenstein zone (conditions for which the distance methods are almost certainly consistent), the number of errors is much higher than likelihood. In fact the distance methods are performing worse than weighted parsimony methods under most conditions (Figure 3.24); the exceptions are in the cases of moderate long-branch attraction. Neither method does as well or better than the other over the entire parameter space studied. This makes deciding between the two difficult. Judged on the size of parameter space in which parsimony performs better and the fact that weighted parsimony gets the true tree right significantly more times (based on performance over the entire parameter space), parsimony appears more robust and/or powerful than distance methods.

In Huelsenbeck's (1995a) study of inference methods on similar trees when data were generated under the simple Kimura two-parameter and Jukes-Cantor models, the performance of distance methods was almost identical to that of maximum likelihood methods. Huelsenbeck concluded that maximum likelihood techniques were superior, but the difference between methods was clearly much more similar than in the simulations which I have done. Like maximum likelihood techniques, distance approaches are consistent when the assumptions of the models are met. My results indicate that there can be a surprisingly large difference in robustness. Yang (1994) and Huelsenbeck (1995b) also report simulations in which maximum likelihood was more robust to ignoring important parameters, such as base frequency bias or a bias in favor of transitions, than methods which relied on corrected distances only. Those results were compelling but only compared very simple models. My results show that even when using the full GTR distance correction accounting for rate heterogeneity, model violations such as non-independence and non-homogeneity can have a much more dramatic effect on distance corrections than maximum likelihood implementations of the model.

The cause of the difference in performance between maximum likelihood and distance methods probably arises from two sources of information. When producing a corrected distance matrix, the characters for all of the taxa but two are ignored. Ancestral sequences are not inferred, so character conflict is not detected and explicitly dealt with. It is common to try to minimize the effect of distance errors by downweighting problematic distances (long distance estimates), but using reconstructed ancestors to detect disagreement in pairwise distance is not presently done. Another advantage that maximum likelihood has over distance methods, is that a character's evolution over the entire phylogeny can be used to inform the algorithm about the rate of a character. Distance methods can accommodate rate heterogeneity by adjusting the expected sequence divergence for any given branch length, but when two sequences differ at a site, no attempt is made to estimate that character's rate and use the rate to inform the distance estimate (this is done implicitly in maximum likelihood implementations of gamma rate heterogeneity – the rate is not inferred as a parameter, but for a character that is evolving very quickly for most of the phylogeny, the likelihood term from the fast rate category will dominate the likelihood of that character).

Amino Acid – Parsimony

The debate over which level of analysis is most appropriate for coding DNA sequences has been hard to study. Clearly there can be useful information in synonymous changes, particularly when closely related taxa are being examined. Despite the information lost in translating data into amino acid sequences, there are arguments for doing so. If non-synonymous changes occur more slowly, they may be less susceptible to homoplasy, and translating sequences into amino acids may significantly increase the signal to noise ratio in the data. A similar line of argument stresses that the large number of states for amino acids may make convergence rarer (although given that not all amino acid pairs are one mutational step apart, it is not reasonable to assume a Cavender-Farris style model of any state changing to any other state at the same rate, and the structure imposed by the code undoubtedly increases the chance of amino acid level homoplasy). If general rules about amino acid evolution can be inferred, then models of amino acid evolution may be better equipped to discriminate homoplasy from true signal.

Parsimony analyses of amino acid sequence were greatly improved through the use of stepmatrices (Figures 3.16 – 3.18 versus Figure 3.15), just as nucleotide-level parsimony analyses were improved through the use of stepmatrices. The PAM1 based matrix resulted in the best performance followed by the matrix based on the number of mutational steps between amino acids (the ProtPars matrix) and the PAM250 matrix. The PAM matrices were derived from comparisons of nuclear-encoded proteins, many of which were globular cytosolic enzymes. Despite the fact that cytochrome *b* is a mitochondrial-encoded transmembrane protein, the PAM weights of the amino acid replacements are more informative than simply considering the minimum number of mutations required to change from one amino acid residue to another. The relatively short branch lengths (branch lengths ranged up to 1.0 expected nucleotide changes per site) may explain why the PAM1 matrix outperformed the PAM250 matrix in these simulations.

Amino Acid – Maximum Likelihood

Maximum likelihood implementing the model of amino acid evolution based on Adachi and Hasegawa’s fit of the reversible model to mammalian mitochondrial sequences was the most effective way to analyze the simulated protein data. This model performed significantly better than any amino acid parsimony weighting scheme. As with the DNA based approaches, the difference is most noticeable in cases for which long-branch attraction is the cause of the incorrect tree being preferred (Figure 3.25); however for the protein sequences there is a larger area of parameter space for which maximum likelihood techniques do noticeably better than parsimony methods. This may reflect the fact that the model of Adachi and Hasegawa is more appropriate for analysis of the cytochrome *b* sequences than the PAM matrix or the fact that there are few informative changes at the amino acid level so explicit modeling of branch lengths is necessary to increase the power of phylogenetic methods.

The proportional model of amino acid evolution (the analog of the nucleotide F81 model), on the other hand was slightly worse than parsimony using the PAM1 matrix. Models of amino acid exchangeability such as the work of Dayhoff or Adachi and Hasegawa, ignore heterogeneity in the forces of evolution across sites. Such heterogeneity is present in the simulated data, but the deviations from the "rules" assumed from the empirical models must be relatively unimportant compared to infomation gained by making generalizations about the patterns of amino acid change because the homogeneous weighting systems dramatically improve the performance of parsimony and likelihood.

Amino Acid vs. DNA Analyses

Whether the data were analyzed under the parsimony criterion or using maximum likelihood, DNA-based methods recovered the true tree more than amino acid approaches. Figures 3.26 and 3.27 compare the performance of the two levels of analysis for each criteria. There is some indication that amino acids are less susceptible to long-branch attraction (amino acid methods do better than their DNA counterparts in the Felsenstein zone trees), but for most problems the loss of information involved in translating the sequences appears to hinder phylogenetic reconstruction. Interestingly, it does not appear that the internal branches are so short that amino acid methods are failing because there is not any true signal; when all of the branches are 0.05 expected nucleotide changes per site, amino acid methods get the tree right in over 95% of the trials. Instead it appears that as other branches lengthen, the noise in the amino acid sequences is able to overwhelm the lower amount of phylogenetic signal more easily. All conclusions of simulation studies are contingent upon whether the data generator produces reasonable data or biases the results in some way; the poorer performance of amino acid methods could result from the selection of cytochrome *b* as the choice of molecules to base the Halpern/Bruno model on. Cytochrome *b* is quite conservative at the amino acid level, so it, or models based upon it, might be expected to be poor candidates for amino acid methods. Fitting parameters of the Halpern/Bruno model with maximum likelihood may exacerbate this problem. Maximum likelihood has a tendency to overfit data. In this context an amino acid with a true frequency that is low, but not zero, could be missed in the mammalian sequences that have been collected. The result is that the maximum likelihood estimate of its frequency will probably be zero, and the amino acid will never occur in the simulated data sets, leading to a further reduction in the number of states in an amino acid analysis.

Chapter 4 – Simulations – Sixteen-taxon Tree

Although there is active research into methods based on solving the phylogeny problem by breaking the full data set up into quartets of taxa, most applied systematists are interested in building trees for larger numbers of taxa. This presents a problem for simulations studies, because an exhaustive coverage of tree shapes (in terms of topological shapes and branch lengths) quickly becomes impossible. While four-taxon studies may provide important insights into how well methods discriminate between signal and noise, it is probably unwise to extrapolate from small trees to make general conclusions about the relative merits of methods. To address large problems in phylogenetics I have performed three other simulations experiments focussing on the performance of DNA based tree inference: one on contrived sixteen-taxon trees, one from a tree shape inferred from real sequences, and one using a random process to generate trees.

Contrived trees can be useful because they allow the study of tree shapes that are thought to be difficult; thus differences between methods are easier to detect. Phylogenetic methods have been verified enough from experimental studies, simulations, and first principles, that most systematists would concede that there is a broad range of trees for which most known methods would correctly estimate the true tree. Further study of these trees is probably unwarranted.

One criticism of the shape of four-taxon trees discussed in the previous section, is that the branch lengths are quite extreme in many cases. For all points that are not on the diagonal of the graph of parameter space, it is not possible to root trees and produce a tree that is consistent with the molecular clock. For any rooting position at least one of the terminal branches with the length determined by the two-branch-length parameter, will be equal in age to a terminal whose branch length is determined by the three-branch-length parameter. For the upper left and lower right corners of the parameter space that I examined, this implies a twenty-fold change in rate for sister taxa. While a strict molecular clock provides a poor fit for most genes in relatively divergent taxa, twenty fold differences in rate are probably uncommon.

Another objection to the four-taxon simulations is in the Felsenstein zone there are only two long branches. Long branches in a tree can be a significant source of noise because of stochastic convergence. In the four-taxon case, the noise is concentrated in favor of one particular incorrect topology, but in larger trees there may be many long branches so the noise may not lead to the attraction of any two particular long branches. So the four-taxon trees may be unrealistically difficult.

Description of the Sixteen-taxon Trees

To address these two criticisms, I developed a parameterization of symmetric sixteen-taxon trees. To simplify parameter space, the trees in this simulation are ultrametric (branch lengths are picked so that they are compatible with a molecular clock). The problem can be thought of as inferring the phylogeny of four groups of organisms, each of which has four members sampled. Two of the groups are “old”, meaning that there last common ancestor was close to the root of the tree, and two are young. Two parameters control the branch lengths of the tree, as shown in Figure 4.1. The branch lengths within the young groups and the length of the branch that creates the deepest split in the tree (dividing the tree into two groups of eight) are kept constant for all of the simulations. The total age of young clades are held such that there are 0.03 expected changes per site from the most recent common ancestor of the young clade to the tip. The central branch of the tree is always 0.04 expected changes per site, this is roughly the same amount phylogenetic signal that was present in the shortest internal branch in the four-taxon simulations.

One parameter (along the horizontal axis of Figure 4.1 and the graphs presenting the results later) controls the depth of the root. As this parameter increases, the root moves deeper in time, but the length of the branch leading from the root to the old clades does not change (so the most recent common ancestor of each old clade and the root of the entire tree both become older). Large values of this parameter lead to greater disparity between the age of the young clades and the old groups, creating a mixture of long and short internal branches. To maintain ultrametricity, the speciation times within the old groups are spread out evenly (the time between the most recent common ancestor of the group and the first split in the group equals the time between the first split and the second split and equals the time between the second split and the present). The second parameter does not change the age of the root, but does change the relative ages of the old and young clades by increasing the length of the branch leading to the older clades. As one moves away from the middle of the graphs, the older clades become more similar in age to the younger group. Along the diagonal of the parameter space all four groups are identical in age, and there is no potential for long-branch attraction between any two particular clades.

The sixteen-taxon simulations are more coarse-grained than the four-taxon study. Moving along the horizontal axis the age of the root changes from 0.07 expected changes per site to 0.97 expected changes per site in nine steps of 0.1 changes per site. Similarly as one moves one step away from the horizontal axis, the time of the origin of the old group moves 0.1 expected changes per site closer to the present.

The trees are based on the same properties that make Felsenstein zone trees hard to infer for parsimony and the “Farris” zone trees more difficult for likelihood. In one set of simulations, shown above the horizontal axis, the two young groups are most closely related. In these Farris zone simulations, the two longest internal branches are sister to each other. In this case, noise along these branches may help methods get the right tree “by accident” because convergence along the long branches agrees with the true phylogenetic signal. In the other set of simulations, shown as the lower triangle, each young group is sister to an old clade, so if homoplasy spuriously unites the internal branches, the incorrect grouping will be obtained. In both cases there are other long branches in the tree, specifically the basal branches in the old clades, so homoplasy should not be concentrated on just two branches.

Methods Examined

Four methods of phylogenetic inference were examined: unweighted parsimony, maximum likelihood using the GTR model with rate heterogeneity, neighbor-joining, and minimum evolution (both based on distances corrected using the GTR model with maximum likelihood estimates of parameters). The parsimony and maximum likelihood searches were performed using the SPR branch swapping algorithm from a random-addition-sequence stepwise-addition tree. The minimum evolution searches were also implemented with SPR-branch-swapping, but the starting tree was the neighbor joining tree. Because of the computational expense of likelihood techniques, searches were done with the parameters of the model set to the values inferred from the most parsimonious tree. The values of the rate heterogeneity parameters were used in the distance corrections. In cases in which more than one tree was returned by a method (this occurred a few times under the parsimony criterion), the first tree found was selected for comparison to the model tree.

Results

The sixteen-taxon simulation was designed to test the methods on their relative power for inferring deep structure of the tree. I had anticipated that the four groups (two old clades and two young clades) would be correctly inferred by all methods over most of parameter space, and that differences would emerge in the relative ability of methods to infer correctly the deepest split in the tree. In both the Farris zone trees and Felsenstein zone trees, there are two symmetrically positioned old clades and two young clades. Because of the symmetry I have combined the results of both old groups into one category, and done the same for the young clades. Thus, the results are presented in terms of probability of recovering: the deep split in the tree, the old groups, the young groups, and the internal structure within the young and old clades. The structure within each of the clades was a pectinate subtree, so there are two internal branches to infer: one separating the two sister taxa from the rest of the tree (referred to as the 1-2 branch), and the other separating three taxa within the clade from the rest of the tree (referred to as the 1-2-3 branch).

Table 4.1 presents the performance of each of the four methods examined for each branch in the tree. Methods that appear in the same cell were not significantly different from each other. In cases in which there were significant differences, the methods are sorted with the left columns containing the best methods. Significance was judged using a likelihood-ratio test with the percentage of replicates in which a branch was correctly inferred treated as a binomial probability. The alpha level was adjusted for the 48 comparisons (6 pairwise comparisons of methods on eight types of branches). Because the percentage of times a branch was correctly inferred over the entire parameter space does not give all of the information about performance, I have plotted the performance in each of the 100 different conditions examined. Figures 4.2-4.8 graphically depict the performance of each of the four methods for each of the type of branch in the tree. The same color scheme was used for the sixteen-taxon study as was used in the four-taxon results. To display more clearly the differences between methods over the parameter space, Figures 4.9-4.15 show the relative performance of methods: the number of times one method recovered a clade in a certain simulation condition minus the number of times a contrasting method recovered that clade. In these figures white indicates that the performance was identical for two methods.

A cursory examination of the results reveals that it was not the case that most of the clades, except the deepest split in the tree, were reconstructed by all of the methods. In fact only one grouping (the monophyly of the four taxa in each of the young clades) was recovered with very high accuracy by all methods. The most closely related pair of taxa in the old clade (Old 1-2 branch) were also recovered in the vast majority of cases. Failure to reconstruct the four basic groups in these simulations make interpretation of the results for any particular branch less obvious.

The simulations can still be informative, but it is important to remember that one taxon being out of place on an inferred tree can cause that replicate to fail to recover several branches.

As in the four-taxon simulations, maximum parsimony proved to be a remarkably robust method. In fact it performed best in terms of the total number of branches missed (summed over all replicates and all conditions), or number of times the entire tree was reconstructed. Maximum likelihood was the next best method when judged by the total number of branches, followed by neighbor joining then minimum evolution.

Recent Branches

The internal structure within the young clade was the part of the tree for which parsimony most clearly outperformed the other methods. In almost all replicates, all of the methods correctly recovered each of the young clades. Given the low divergence within the group, it is unlikely that any method incorrectly inferred the unrooted topology of the four taxa within the young clades. It appears that maximum likelihood and distance methods had trouble correctly rooting this subtree, while parsimony almost always recovered both branches within the clade (young 1-2, and young 1-2-3 Figures 4.2, 4.3, 4.9 and 4.10). This performance may be due to parsimony’s bias toward uniting long branches (in this case, the longest branch in the subtree leads to the basal member of the clade and in the true tree this long branch is connected to the long branch uniting the young clade to the rest of the tree).

Intermediate Branches

While parsimony performed better than the distance methods for virtually every branch in the tree, the contrast between parsimony and maximum likelihood is more complex. As mentioned above parsimony recovered more of the recent structure of the tree (the internal structure of the young clade). On branches of intermediate depth (the internal structure of the old clades), the methods were not significantly different based on the entire parameter space. On the old 1-2-3 branch

(Figure 4.6 and 4.13) parsimony appears more sensitive to long branches combined with a large disparity between the ages of the clades (right portion of the graph near the middle of the graph, which is this simulation study's equivalent of the extreme Farris and Felsenstein zones), while likelihood is failing more when the ages of the clades are even. Along the diagonal the old clade is not actually older than the young clade, and likelihood is presumably having difficulty rooting the subtree as discussed above in the context of the young clade.

Deep Divergences

Maximum likelihood is outperforming parsimony on the oldest divergences in the tree (recognizing the old clades as groups, and inferring the deepest split in the tree), but the difference is not enormous (Figures 4.7, 4.8, 4.14 and 4.15). Interestingly maximum likelihood is correctly inferring the deepest divergence in the tree more than parsimony whether the internal structure of the tree mimics the Felsenstein zone or the Farris zone, although no method is reliably inferring deep branches in the extreme parts of parameter space (and the difference between the methods is more pronounced in the Felsenstein zone like trees).

Distance Methods

As was the case for the four-taxon tree simulations, distance methods performed worse than parsimony or likelihood. Unlike the four-taxon study, there is no region of parameter space in which distance methods outperform parsimony. Neighbor joining does slightly better than minimum evolution at recovering every branch in the tree.

Obviously distance methods rely critically on the distance correction that is used. While the GTR model with rate heterogeneity is favored over other single nucleotide models, based on likelihood scores of trees under different models, this model is not necessarily the best for correcting distances. Steel and Penny (2000) have shown that p distances may be preferable to corrected distances when the tree is ultrametric. Discussion of the performance of distance methods using less parameter rich models will be postponed until Chapter 7. In these simulations the performance of distance methods was dramatically improved by using p distances. The results from the GTR model with rate heterogeneity are presented because this is the favored model under standard model-selection criteria, and, entering an analysis, many researchers are unlikely to assume ultrametricity (which happens to be valid in these simulations).

Conclusions

The conclusions of the sixteen-taxon study are quite similar to those of the four-taxon study. Overall performance appears lower, but this is largely the result of at least one branch on the tree (the deep divergence) being very short and surrounded by much longer branches. In essence there were many fewer trees that were easy in the sixteen-taxon simulation. Parsimony seems to be remarkably robust. Maximum likelihood using the standard single-nucleotide models improves reconstruction of deep nodes, but the improvement over parsimony is not enormous. In this simulation there are indications that maximum likelihood may perform worse than unweighted parsimony for some branches. Unfortunately it is not clear whether this performance is due mainly to a failure of maximum likelihood, or a bias in parsimony. Distance methods implementing the same model as maximum likelihood behave notably different and worse than likelihood or parsimony, and neighbor joining outperformed minimum evolution searches over most of the parameter space.

Chapter 5 - Simulations - 228 Angiosperm Tree

In 1996, Hillis reported what is probably the most astonishing result from a phylogenetic simulation study. Using Kimura's two-parameter model with rate heterogeneity (gamma-distributed rates with a shape parameter of 0.5), he simulated data onto a tree taken from a parsimony analysis of 228 sequences from a wide diversity of angiosperms. Surprisingly, very simple methods, such as unweighted parsimony stepwise addition searches performed quite well. The full tree was recovered without error with just 5000 bases of data (after branch swapping), despite the fact that the model tree had great heterogeneity in branch lengths, including some very short internal branches. To demonstrate that this result was robust, he lengthened all of the branches on the tree to ten times their length as inferred on the real data and repeated the simulations. The longer tree was actually easier to infer, based on comparing the number of branches missed as a function of the number of bases simulated. Hendy and Penny (1989) had suggested, from theoretical studies and analysis of small trees, that the addition of taxa may make some phylogenetic problems easier to analyze. Later simulations studies have also shown that addition of taxa can help avoid problems associated with long-branch attraction (Graybeal 1998, but also see Poe and Swofford 1999), but the high profile of Hillis' study and its large impact on the field make it the most obvious choice of a simulation study to replicate.

Hillis' results were surprising in light of the large number of four-taxon simulation studies (Huelsenbeck and Hillis 1993, Huelsenbeck 1995). While phylogenetic methods performed well over a large range of parameter space, there were combinations of branch lengths that made it very difficult for parsimony (and other mehods) to reconstruct a branch correctly. Merely extrapolating from the number of bases that it took to infer a single internal branch with high probability seemed to imply that a huge amount of data would be needed to reconstruct the 225 internal branches of the angiosperm tree. If the probability of reconstructing each branch were 0.95, and the branches were treated as independent, then the full tree would only be recovered one time in 100,000 trials. Even if the probability of getting each single branch were 0.99, there would only be a 10% chance of getting all 225 branches correct. Reconciling Hillis' results with these calculations can be done by realizing that it is not correct to treat the probabilities of reconstructing different branches on the same tree as independent (correctly inferring one branch greatly increases the chance of getting a nearby branch too), and that one character can serve as a synapomorphy for several branches on a large tree.

Traditionally systematists have placed a high value on homoplasy-free characters and finding objective ways of coding character states so that the characters are likely to be free of homoplasy. DNA sequence characters are prone to frequent changes and cannot be coded in such a way that homoplasy is avoided. Hillis' study showed that it was possible to reconstruct a large trees from relatively noisy characters. In fact, if one simulates homoplasy-free binary characters on the angiosperm tree, the number of characters needed to recover the tree would be larger than in his simulation in which homplasy is allowed (data not shown). Whether these results are robust to the use of a more complex model for data generation is an interesting and open question. Gamma-distributed rate heterogeneity in Hillis' original work leads to some characters evolving with a high rate and, hence, a considerable amount of homoplasy. This type of simulation also produces a few characters that change very slowly and tend to have very little convergence or parallelism. While Hillis showed that his conclusions were not strongly sensitive to the particular branch lengths used, it is not evident if the results were contingent on the pattern in which homoplasy accumulates (i.e., the model of evolution).

I have replicated Hillis' study by simulating the data based on the parameterized version of the HB model. I simulated data on the empirical branch lengths (which I will refer to as the short tree) as well as a tree with branches ten times longer than the empirically based estimates (the long tree). To vary the amount of data available to the inference methods, I concatenated independently simulated data matrices. Hillis' study covered up to 5000 bases of simulated sequences. I studied performance on one, two, four, eight, and sixteen copies of the simulated data. Thus the lengths of the sequences covered varied from 1128 to 18,048 bases. For each set of conditions 100 replicate data sets were simulated and analyzed.

Hillis found that unweighted parsimony performed slightly better than neighbor joining. I tested the performance of unweighted parsimony, neighbor joining, and minimum evolution. For the distance methods, distances were corrected using the GTR model with estimates of the gamma shape parameter and the proportion of Invariant sites provided by maximum likelihood estimates from one simulated data set (because maximum likelihood estimates from each replicate would be too computationally intensive). Parsimony was investigated under three search strategies: one stepwise addition search using a random addition sequence, one SPR search keeping one tree from a random addition sequence stepwise addition tree, and one SPR search starting from the model tree. Minimum evolution was implemented as one SPR search from the neighbor joining tree and one SPR search starting from the true tree.

Starting searches from the true tree is a heuristic tool to help determine whether branches of the true tree are missed because of an inefficient search strategy (poor solution to the optimality criterion) or whether the cause of failure is a result of the optimality criterion favoring an incorrect topology (this strategy was suggested to me by T. Warnow, personal communication). Such searches can be informative but difficult to interpret correctly. They represent a first order approximation of an upper bound on how well the optimality criterion might perform if the search could be done exhaustively. They do not represent how the criterion would perform in an exhaustive search because it is possible that a tree exists that is topologically less similar to the true tree but has a better score under the optimality criterion. Searches from the true tree do not provide a strict upper bound on the performance of a method because it is sometimes possible to find a topology which has a better score and is closer to the true tree (in fact this happened three times in parsimony searches in the current study). The utility of starting a search from the true tree is to help assess how reasonable it is to expect a method to perform better if more intensive searching were performed.

It is interesting to note that, in Hillis' study, simulations on the larger of the two trees resulted in sequences that look biologically implausible. He states (1996): "Under these conditions, the average character is changing 23.6 times across the tree, and because of rate heterogeneity among sites, some characters change many more times. At these high rates of evolution, many of the terminal sequences are so dissimilar that no biologist would recognize them as homologous." The implication seems to be that the sequences are much more divergent than sequences regularly used in phylogenetic analysis. The branch lengths of the tree under this model result in a root to furthest tip distance of about 0.9 expected changes per site. While this is undoubtedly a long branch, this level of divergence is displayed within mammals for the cytochrome *b* gene. In fact, based on the branch lengths inferred under the HB model, within Murid rodents there are taxa that exceed this level of divergence. When data are simulated on the long tree under the HB model, the result is sequences whose homology no biologist would question. Plausible looking sequences under these branch lengths result from a more biologically realistic spatial arrangement of conserved sites and the lack of any sites evolving at extremely high rates.

The branch lengths for the long tree simulation are clearly within the realm of plausible sequence evolution for genes that have been used in phylogenetics. As mentioned earlier, cytochrome *b* evolves quickly (at the third base postions), and many researchers (myself included) would question its usefulness for deep relationships. I do not wish to argue that sequences with these rates of evolution should be used for reconstruction of trees, but, by using a more complex simulator, it is apparent that divergent sequences do not always stand out as random or hopelessly distant.

Results

Figures 5.1 - 5.6 show the results for each method. In each case the results for the tree with empirically based branch lengths are shown in blue and the results from the long tree are shown in red. The mean of the 100 replicates is shown as a solid line; the dashed lines indicate 95% confidence limits. The Y axis is the percentage of the 225 internal branches in the model tree that were recovered by the method. The x-axis displays the number of copies of the cytochrome *b* modeled genes concatenated together.

Discussion of Parsimony Results

Stepwise Addition

The performance of stepwise addition on the smaller of the two trees is very similar to the results given by Hillis (see Figure 5.1). The tree is not recovered quite as accurately, despite the much longer sequence lengths used in this study, but clearly stepwise addition is doing a remarkably good job at inferring the tree, attaining a mean reconstruction success of over 95% with four copies of the simulated genes and over 98% when sixteen genes are simulated. For the longer tree, the results for stepwise addition searches are qualitatively different from those seen when data is simulated under a simple model like K2P. The performance of the algorithm is dramatically worse than the performance on the shorter tree. With one copy of the gene, on average fewer than 63% of the branches were estimated correctly. Even with the full 18,000 bases of simulated datam, the performance only improves to about 81% of the tree being correctly inferred.

Parsimony Searches

On the smaller tree, branch swapping does not dramatically affect the performance of parsimony but helps a small amount (Figure 5.2). Most of the branches that were missed in the stepwise addition search are not due to the use of a quick approximate solution to the parsimony optimization problem, instead they represent examples of the optimality criterion favoring a tree that does not contain a branch found in the true tree. More thorough searching noticeably improves the performance for short sequence lengths, but for simulations with eight or more gene sequences, the differences between stepwise addition and a search with branch swapping are slight. Given enough data the stepwise addition algorithm is quite efficient at producing a good estimate of the most parsimonious tree for these simulations. Searches started from the true tree performed slightly better than starting from a random addition sequence stepwise addition tree (Figure 5.3). The improvement from starting a search at the true tree was most noticeable for short sequence lengths.

Branch swapping had a profound effect on the performance of parsimony when the data were simulated on the larger tree. In this case stepwise addition searches performed poorly, but after more thorough searches the performance on the long tree was only slightly worse than the performance on the short tree. With 16 genes simulated, about 97% of the true tree was recovered. Once again searches from the true tree did only slightly better than searches from a stepwise addition tree. On real data sets for this number of taxa the preferred topology usually changes substantially when branch swapping is done, so it is not surprising that there is a dramatic difference between the stepwise addition tree and the tree after an SPR search. For simulations with one copy of the gene branch swapping resulted in a tree with, on average, 44 more correct branches. For longer sequences, the effect was smaller, but still substantial (about 36 more true clades were identified by the SPR search). Hillis' conclusion that a great amount of phylogenetic signal can be recovered from rapidly evolving sequences (if the tree is densely sampled) is robust to simulations based on much more complex models of evolution.

Neighbor Joining

In Hillis' study on the smaller tree, neighbor joining performed slightly worse than stepwise addition. Hillis did not discuss the performance of distance methods on the long tree, but I have replicated his simulation and verified that neighbor joining performs well on the long tree with 5000 simulated bases (data not shown). Qualitatively both neighbor joining and stepwise addition exhibited the same pattern: remarkably high accuracy with reasonably long sequences. Using the HB model to generate sequences, the two methods gave similar results (Figure 5.4). On the shorter tree, neighbor joining appears to be slightly outperforming stepwise addition (better mean performance for all of the sizes of data sets examined). On the longer tree, neighbor joining appears to be performing slightly worse than stepwise addition, but neither method is doing well.

Minimum Evolution

Minimum evolution searches were performed by SPR branch swapping from the neighbor joining tree. For the unweighted parsimony criterion, branch swapping slightly improved performance on the small tree. This was not the case with minimum evolution. Minimum evolution searches resulted in trees that were no better than the neighbor joining tree in terms of topological distance to the true tree (Figure 5.4). This result is not surprising given that neighbor joining did quite well on the trees, so there was little room for improvement. However, on the longer trees the quick heuristics performed poorly, and yet estimates of the minimum evolution tree were actually worse than the neighbor joining tree in most cases. The minimum evolution criterion also performed poorly when the searches were started from the true tree (Figure 5.6), in fact the final trees from these searches were almost identical in terms of distance from the true tree. This indicates that the poor performance of minimum evolution is unlikely to be the result of cursory searches for the minimum evolution tree.

These results are surprising in light of the common tendency to view neighbor joining as simply a starting point for more intensive searches. In these simulations neighbor-joining seems to be detecting more signal than the minimum evolution criterion, and searching hurts performance more than it helps. Also remarkable is the dramatic difference between the distance-based approaches and parsimony on the longer tree. The four-taxon studies indicate (and it has been noted elsewhere) that distance methods can perform poorly when the evolutionary distances are large, but results from simulations studies done with simple models of evolution have not shown as great a difference between distance and parsimony methods as observed here. Nei and Kumar (2000) point out that all methods perform similarly when divergence is low (less than 0.025 changes per site), which is undoubtedly true. They assert that pairwise divergences between taxa of greater than 1.0 are "biologically unreasonable," presumably meaning that phylogenetic analyses should not be performed on such sequences. As mentioned above, the divergences considered in this set of simulations is within the divergence of one group of rodents. The long tree simulations seem to agree with Nei and Kumar's assessment that phylogenetic analysis using distance approaches is unreliable for this level of divergence, but parsimony is still performing remarkably well on this tree indicating that phylogenetic analysis in general is not hopeless. Of course these conclusions are based on GTR distances with two types of rate heterogeneity. While there is enough information in the data to support this model in the maximum likelihood framework, it is possible that other distance corrections (for instance a method based on non-synonymous changes only) would improve distance analyses.

Angiosperm Tree Final Thoughts

It is tempting to claim that these simulations show that parsimony will be a powerful tool for inferring phylogenies of large numbers of taxa even when the data are generated by a very complex process and that these simulations provide striking examples of a performance gap between parsimony and the commonly used nucleotide distance methods. Other than the caveats that are appropriate for all simulations using this model of sequence evolution, it is also important to consider possible biases that arise from the tree shape. The topology of the tree and branch lengths were inferred using parsimony. This may constitute a substantial bias in favor of parsimony.

Parsimony seems to be more susceptible to long-branch attraction than other methods, but this does not appear to be a large problem on this tree. This could indicate that the conditions that lead to long-branch attraction are rare in real trees, or it could be the result of using parsimony to infer the original tree. If there had been a pair of long branches separated by a short internal branch in the real tree, it is plausible that the branch attraction occurred in the analysis of the real data, so that the two long branches were united in the model tree used in this study. Rannala *et al.* (1998) inferred a different tree from the same data using UPGMA and used this tree as a basis for simulations; parsimony performed well on this topology but required much more data to reach high accuracy than it did on the model tree inferred from parsimony.

Similarly the branch lengths estimates for this model tree were based on parsimony inferences of the number of changes. Parsimony will underestimate the total number of changes on the tree, and may tend to push changes from terminals to the internal branches (there is no parsimony signal for a polytomy; so if two sister branches happen to display a parallelism, the change will be inferred to have occurred on the internal branch uniting them). This method of estimating branch lengths also puts a limit on the smallest non-zero branch length on the tree (essentially the smallest length is one divided by the total length of the 18S sequences).

These imperfections and biases probably have little effect on Hillis' main conclusion that large trees with remarkably short branch lengths can be inferred from reasonable sequence lengths. The importance of Hillis' paper and its conclusions made it an obvious choice of a study to replicate using the Halpern/Bruno model, however, because I am also interested in comparing the performance of parsimony and distance methods on large trees, it is more important for me to avoid any possible bias in favor of parsimony.   Thus I have conducted a final set of simulation studies using randomly generated trees.

Chapter 6 – Simulations – Randomly Generated Trees

The study of tree shapes is an exciting but difficult area of phylogenetics. A fuller understanding of the shape (topology and branch lengths) of trees would provide insight into the processes of speciation, extinction, and molecular evolution (particularly the rate of molecular evolution and how it changes over time). The field is still in its infancy because there are very imposing barriers to drawing robust conclusions about biological tree shapes. First, there is a very large amount of variance in the tree shape produced by any reasonable model of speciation and extinction; the probability density functions of these models are fairly flat over topology space. This means that it is difficult to decipher which model best fits the real data unless a large number of trees are examined.

Another impediment to the field is the fact that we cannot witness the tree without error, instead we must infer it from analysis of the species. Inference of the tree can be a source of error and bias. The shapes of inferred trees will be affected by the taxon sampling performed by the systematist. Usually not all of the extant species in a group are included in a phylogenetic study, and the included taxa are not a random sample of the whole group. Geography, previous taxonomy, and levels of divergence of interesting traits can affect which taxa are used and cause the data set to be non-random. Even if the sampling were complete (or random) topological biases in the algorithms used to construct phylogenies can lead to a biased set of trees (Huelsenbeck and Kirkpatrick, 1996). Given that there is no consensus on the most appropriate distribution of tree shapes or the best models to use for tree generation, I have generated random tree shapes under a very simple pure-birth process with complete sampling (see Rannala *et al.* 1997 for a discussion of the effect of incomplete sampling on the shapes of birth-death trees and the difficulty of phylogenetic inference on these trees). The tree generation process that I used is most similar to that of Bininda-Emonds *et al.* (2000)

Tree Generation

Trees were created using the assumptions of the Yule, or pure-birth, process (Yule, 1924). These trees are ultrametric, meaning that the root tip to branch length is the same for all taxa. Because the constancy of the molecular clock is doubtful for moderately divergent taxa, the branch lengths were modified away from ultrametricity. There are three parameters relevant to the process of tree generation used in this experiment: the root to tip age of the tree, the mutation rate, and the variation in the rate of molecular evolution. The trees are produced in a four step process: creation of a Yule tree, scaling of the tip to root age, changing branches by allowing the relative mutation rate to vary, and scaling the tree by a mean mutation rate.

Yule trees can be generated easily with the following algorithm:

1. Start the tree with two lineages sharing a common ancestor with a branch length of zero from the ancestor.

2. Calculate the time until the next speciation event by assuming the waiting time follows an exponential distribution with a mean of the reciprocal of the number of lineages.

3. Add the waiting time to the lengths of all of the terminal branches.

4. Randomly pick a lineage to speciate (this tip becomes a parental node and two sister lineages are created with branch lengths of zero from this parent node).

5. Repeat 2 –4 until the desired number of taxa are obtained.

When the process is completed the most recent two species will be separated by branches of length zero. The time until one more speciation event was calculated, and a new branch length was calculated as a uniform random variable over the range zero to this next speciation time. The new branch length was added to all of the terminal branches.

The Yule process generates ultrametric trees, but the tip to root branch length is not the same for all of the trees generated. Because the scale of the problem can have a profound impact on the efficiency of phylogenetic analyses, I scaled all of the Yule trees in a given simulation set to have the same tip to root length. Thus the speciation model was provided a variety of topologies and relative branch lengths, but the overall "age" of the simulated clades was kept constant.

The molecular clock is not universally acceptable, and the degree of ultrametricity can severely affect the performance of phylogenetic methods. So deviations from a clock can be studied in a parametric way, I allowed the relative mutation rate to vary over the tree using a parameter, r, to quantify the rate of evolution of the rate of evolution (following Thorne *et al.* 1998). The rate of evolution at the end of a branch is a found by multiplying the rate at the beginning of the branch by a lognormal variate. This number is calculated by taking e (the base of the natural logarithm) to a power which is a random number drawn from a normal distribution with mean of 0 and a variance equal to the branch length multiplied by r. Thus the exponent is a random variable with a mean of zero, but high variance when branches are long or when r is high. Presumably the rate of molecular evolution cannot be zero, and clearly it cannot be infinite. To constrain the rate of evolution, a ceiling of ten times the original rate and a floor of one tenth the original rate of evolution were enforced. This allowed a maximum rate difference of 100 fold on different branches of any one tree. Once the rate at the end of a branch was calculated the relative rate of evolution for the whole branch was calculated by averaging the rate at the beginning of the branch and at the end of the branch. In cases in which the rate of evolution exceeded the maximum or fell below the minimum, the evolution of the mutation rate was reflected back into space of legal parameter values.

After branches were modified to allow for variation in rate, the tree was rescaled by multiplying all of the branches so that the total length of the tree (the sum of the branch lengths) was identical before and after accomodating lineage specific rate variation. This was done so that changing r would not change the size of the tree (to decrease overlap in the effects of the r and the tree scaling parameters). Thus r affects the distribution of mutations across the tree as a whole, but does not affect the total number of mutations. Note that, older trees will have more rate changes, because the amount of change in the rate of evoluton across a branch is determined by r and the original length of the branch.

The final modification of the trees allowed the specification of a mutation rate at the beginning of the tree. Given the branch lengths and the relative rates of evolution on each branch, specifying the mutation rate amounts to simply multiplying each branch by a constant factor. The mutation rate and age of the root determine have a strong effect on the total length of the tree and hence determine the how long the simulated trees are.

Simulation Details

With three continuous parameters governing tree construction (in addition to the number of taxa), it was not plausible to investigate the full range of parameter space. The main goal of the random tree simulations was to investigate trees with a large number of taxa to verify that the general conclusions from the smaller simulations were robust. Random trees were needed to ascertain whether the bias in favor of parsimony on the angiosperm tree was leading to the disparity between the performance of parsimony and distance methods. Given these goals I chose a strategy of sweeping over a relatively large portion of the space of the tree generation parameters. r was set to three values 0.1, 1.0, and 10. To simplify the parameter space, both of the tree scaling parameters, the age of the root and the mutation rate were kept equal to each other. Simulations were performed with these parameters set to 0.5, 0.6, 0.7, 0.8, 0.9, and 1.0. Because both of the parameters were changed simultaneously, the effect on the scale of the trees is more or less the square of the parameters. Hence trees with these parameters set to 0.5 are on average four times shorter than trees with the parameters set to 1.0. This strategy ensured a broad range of tree depths were covered. Note that pairwise divergences between taxa across the root of the tree are expected to be around two times the age of the tree times the mutation rate, so when the scaling parameters were set to 0.5 the largest pairwise divergences were 0.5 (in terms of expected changes per site). When the scaling parameters were set to 1.0, the largest divergences were 2.0. Deviations from the molecular clock (increasing r) can increase these maximum divergences. These trees are divergent by most researchers' criteria (though not dramatically longer than the long version of the angiosperm tree investigated by Hillis). Because it is difficult to visualize the effects of these parameters, I have included six figures (6.1-6.6) depicting twelve simulated trees from the most extreme parameter values used. The scale of the trees is not constant; the figures are included to help understand the wide range of tree shapes that were studied.

Trees of 50 and 100 taxa were generated using the procedure described above. Given the three values of r and the six values for the scaling parameters, there were 18 conditions examined for each number of taxa, for a total of 36 simulations conditions. For each set of parameters 100 trees were generated and one data set was produced for each tree (as opposed to the previous studies in which the tree was kept constant and replicate data sets were produced). Each of the data sets was analyzed using two search strategies for unweighted parsimony: ten random addition sequence stepwise addition searches, and ten searches using SPR branch swapping from a random addition sequence stepwise addition tree. During the searches no more than 100 trees were retained. At the end of each search one of the most parsimonious trees was picked for comparison to the true tree. Neighbor joining and minimum evolution searches were performed using the GTR distances and rate heterogeneity, using the same values of the gamma shape parameter (1.99) and proportion of Invariant sites (0.475) that were used in the angiosperm tree (these values were estimated from one simulated data set on that tree, and represent fairly robust estimates of the rate heterogeneity values when a large data set is fit to data from the HB model). To mimic the searches that were done using the parsimony criterion, the minimum evolution searches were started from stepwise addition trees with ten searches performed from different random addition sequences.

Simulation Results

Figures 6.7-6.12 show the mean performance as measured by the percentage of internal branches in the true tree that were recovered by each method. Each graph shows the results for all four inference methods for one setting of r and all six values of the scaling parameters. Performance of the parsimony methods is shown in blue; distance methods are shown in red. The fast heuristics (stepwise addition and neighbor joining) are plotted as dashed lines. The best possible performance is also plotted as a black line. These data come from estimating the percentage of internal branches that have at least one substitution occurring on them during the simulation. Without a mutation along the branch, the model tree was effectively a polytomy for that branch, rather than remove such branches from consideration (which makes it difficult to see how many branches of this type occurred), I have simply plotted the performance of an imaginary method that recovers every internal branch which is long enough to have had a mutation on it.

Over the entire range of tree shapes examined there is a wide variation in performance. The best method on average recovers 85% of the branches on the shortest and most ultrametric trees, while for the longest trees, the minimum evolution is only recovering 24% of the correct branches. In general, low values of r and low values of the scaling parameters made the trees easier for all methods.

Perhaps unsurprisingly, given the literature on the importance thorough taxon sampling (Hillis 1996, Graybeal 1998), the percentage of the tree that is recovered by parsimony is relatively unaffected by the number of taxa in the analysis. Because all of the taxa generated by the Yule trees were included in the analysis, it is not correct to treat the 50-taxon data sets as if they were comprised of the same trees as the 100-

taxon trees, but with only half of the taxa included. Instead under the parameters used in this study, these simulations investigate whether or not a group that has diversified into 100 species is more difficult to infer than a group that is of the same age but with half as many species. Adding more taxa means that there are more speciation events in the same amount of time, so internal branches are shorter in the 100-taxon trees (compared to the 50-taxon trees with the same parameter values). Apparently, this effect is balanced out by the improvement in the ability to infer dense trees, such that there is essentially no difference in performance under the parsimony criterion. For extreme branch lengths the trees with r>0.1, distance estimates are missing a higher percentage of branches on the 100-taxon trees than on the 50-taxon trees.

The angiosperm simulations (on the longer tree) provide examples in which performing branch swapping is vital to obtaining reliable trees. Interestingly the performance of parsimony after branch swapping was only slightly better than stepwise addition on the randomly generated trees. It is possible that the larger number of taxa in the angiosperm-based tree is the factor that causes stepwise addition to find a tree far from the most parsimonious tree, however in the random tree simulations there is no indication that branch swapping is more important for the 100-taxon trees compared to the 50-taxon trees.

Bininda-Emonds *et al.* (2000) studied similar trees to infer the rates of convergence of parsimony (how much data were required to recover the tree). They analyzed the ease of inference of branches in the tree and found that the depth of the branch was a crucial determinant of how much data were need to reconstruct it. Shallow branches (such as those uniting sister species) were quite easy to reconstruct. The data were simulated under a K2P model with gamma-distributed rate heterogeneity. Replication of that study with a more complex model of data generation would be very worthwhile. Unfortunately, the data sets they examined involved much larger trees (thousands of taxa), required more sequence data than the

HB model data provide (without resorting to concatenating multiple simulations), and a considerable investment of computational resources. In addition to clade depth, random tree simulations could provide information on how a variety of factors such as the rate of evolution along a branch and branch lengths of neighboring branches affect the probability of reconstructing a node.

Of central importance for this dissertation is whether or not the patterns seen in the previous simulations based on the HB model are artifacts of the types of trees examined or whether they represent robust conclusions about the performance of methods on data simulated under the HB model. Fortunately the most striking conclusions from the previous simulations are also born out by the study of randomly generated trees. Parsimony is outperforming distance methods. This is most noticeable on long trees; neighbor joining is consistently producing more accurate estimates than minimum evolution. In fact, these patterns are true of each 36 combinations of parameters used in the study (as can be seen from the graphs in which the order is the same and the lines never cross each other).

Chapter 7 – Conclusions, Implications and Extensions

In the previous chapters I have described the performance of phylogenetic methods on simulated data that are much more complex than the data used in any previous simulation study. The focus has been on fairly difficult tree shapes and sizes. In general the results indicate that the currently used methods are not as reliable as they seem based on simulations on simple models, but they do perform remarkably well.

Coming into this study, there was a concern that much of the apparent advantage that model-based approaches had over simple approaches like parsimony may be an artifact of data generation in simulations being much too simple. This study supplies evidence that these fears were warranted to some degree. When the assumptions of the model are severely violated maximum likelihood fails to recover difficult trees shapes, such as the extreme Felsenstein zone trees, that it can reliably reconstruct when the data conform to the assumptions of the models of evolution.

This does not imply that model-based approaches should be avoided. Maximum likelihood based on simple single nucleotide models of evolution still outperformed all other methods on the four-taxon trees. The main difference between the results of this study and simpler simulations was that the range of tree shapes in which maximum likelihood outperformed parsimony was smaller (largely because both methods were failing on very difficult trees). On the sixteen taxa trees, parsimony outperformed likelihood on two of the branches (and this may be the result of bias) while likelihood did better on the two deep branches.

Because these simulations were based on only one gene, they do not represent a satisfactory method of generating sequences long enough to thoroughly test recently developed, parameter-rich approaches to phylogenetic inference. Preliminary results indicate that amino acid models outperform amino acid parsimony, but are not as reliable as analyzing data at the DNA level. When fit to data from more genes, the HB model should provide an excellent testing ground for the very divergent approaches that have been suggested: codon models (Muse and Gaut, 1994; Goldman and Yang, 1994), a wide range of amino acid models, and covarion models (Galtier, 2000). If the HB model parameters are allowed to vary over the tree, even non-stationary models (Galtier and Guoy,1997) could be considered.

One of the most robust conclusions of this study was that distance methods were performing poorly in comparison to character based approaches. Undoubtedly this is, to some degree, an artifact of this study focussing on trees with long branches. Nevertheless the fact that other methods (in some cases methods assuming the same model of sequence evolution) were performing significantly better indicates that distance approaches have relatively low power and/or they are very sensitive to violation of the assumptions of the model used to correct distances. Clearly the terminology "distance methods" is a bit misleading. The neighbor joining method or minimum evolution criterion are provably consistent, and they even have proven bounds on their error (given estimates of the evolutionary distances at which the error is below a certain threshold, they are guaranteed to recover the true tree). Thus failure to reconstruct the correct tree is attributable to error in the input distance matrix, not the methods *per se*. In the simulations described in the previous chapters, the model of sequence evolution used to correct distances was chosen using maximum likelihood techniques. Some authors (e.g. Nei and Kumar, 2000) have suggested that distance estimates from simpler models may outperform more complex models, even when the latter are justified by improved fit to the data. For three of the simulation studies I have examined the performance of simple distance corrections.

Simple Distance Corrections

Model-based inference of a complex parameter, such as a phylogeny, is never based on the assumption that the model perfectly describes reality. Burnham and Anderson (1998) provide a good discussion of how model-based inference should be based on choosing a model that is complex enough to avoid biases associated with ignoring important parameters but simple enough to avoid excessive variance from overfitting the data to too many parameters. In this light, it is not surprising that simple distance corrections sometimes outperform a more complex model-based approach. Model-based distance corrections involve estimating parameters from the observed matrix of pairwise sequence differences and then finding a branch length that minimizes the difference between a model's expectation of the distance matrix and the observed data. As the number of parameters grows and as the sequences become more divergent, the variance of the estimated corrected distance grows.

In this study I selected a model based on a comparison of tree-based likelihoods of the data under different models. This seemed reasonable because it meant that maximum likelihood and distance techniques were making similar assumptions, and parameters which were not justified were avoided. Comparing distance methods assuming equal base frequencies and only two substitution types (K2P model) to maximum likelihood under the GTR when there is clear evidence of unequal base frequencies seems unfair to distance methods. Using the tree-based likelihood of the data set may lead to overfitting of the model of distance correction. This is because distance methods do not use as much information from the data as likelihood approaches. Because much of the information is lost when a matrix of pairwise distances is constructed from character data, it is reasonable to assume that the power to estimate model parameters accurately is reduced. Unfortunately the state of the field in terms of model selection for distance corrections appears to be quite unsettled. As opposed to objective statistically based criteria, proponents of distance methods provide either no suggestions or loose guidelines such as:

> "When 0.05<d<1.0 and the number of nucleotides examined is large, use the Jukes-Cantor distance unless the transition/transversion ratio (R) is high, say, R>5. When this ratio is high and the number of nucleotides (*n*) is very large use the Kimura distance or the gamma distance. However, when the number of sequence (sic) is large and *n* is relatively

> small, the p distance often gives better results unless the evolutionary rate varies extensively with evolutionary lineage" (Nei and Kumar, 2000, page 112).

A reasonable methodology might be to use the likelihoods of the pairwise sequence comparisons (instead of tree-based likelihoods) in the standard process of model selection using the likelihood-ratio test statistic or Akaike Information Criterion. This ignores the fact that all of the pairwise distances are not independent data points, and it also means that a model with rate heterogeneity across sites can never be preferred (rate heterogeneity parameters are unidentifiable from pairwise distances because an observed divergence can be explained equally well by a short branch length and no rate heterogeneity or long branch length and strong rate heterogeneity. Another option would be to use a goodness of fit criterion, such as least squares, to test the fit of the distance on the inferred tree and choose a model of distance correction only if it produces a tree with significantly better fit. This would involve several tree searches, but, given the speed of neighbor joining, this is not a serious drawback.

Because there is not a widely agreed upon technique for choosing a distance correction, and to verify that the relatively poor performance of distance methods was not the result of my choice of model corrections, I investigated the performance of several simple distance corrections on three of the simulations described in previous chapters.

Four-taxon Trees

Rate heterogeneity, particularly a model with invariant sites, can greatly increase the variance of a distance correction because in essence a portion of the data is simply removed from consideration. In the data presented earlier the distance corrections employed large amounts of rate heterogeneity; when invariant sites was the preferred form of rate heterogeneity, the estimate of the percentage of unchanging sites was around 50%. The first simplified model of distance correction I examined was a GTR distance with invariant sites, but the proportion of invariant sites was set to 0.25. I also tried several corrections (GTR, HKY, K2P, and JC) with no rate heterogeneity at all as well as uncorrected distances. Judging by the total number of replicates in which the true tree was inferred (i.e. summing over all of the branch lengths examined), performance was improved substantially by the use of simpler distance corrections. GTR with the preferred values of rate heterogeneity parameters (based on maximum likelihood) recovered the tree in 74% of the four-taxon simulations. Lowering the amount of rate heterogeneity improved the performance to 79%, and ignoring rate heterogeneity resulted in the true tree being found in 81% of the replicates. HKY and K2P distance corrections resulted in 82% accuracy, and, with the JC correction, the performance of minimum evolution peaked at 83% success. The use of simple p-distances resulted in a slight decrease to 82% recovery of the correct tree.

Even under the most favorable conditions, minimum evolution is performing significantly worse than weighted parsimony (which succeeded in 88% of the replicates), but it did surpass unweighted parsimony. The improvement seems to be due to decreased variance (as opposed to an increase in the size of parameter space in which minimum evolution performs well). This conclusion is based on the fact that improvement seems to come in regions in which distance was already doing well. In fact, performance in the Felsenstein zone is hurt by using only the JC correction (as shown in Figure 7.1). The results actually present an interesting problem for a researcher: overall the distance methods are working much better with simple distance corrections on these trees, but there is actually no region in which distance methods are outperforming weighted parsimony (see Figure 7.2). This is in contrast to the case of minimum evolution employing the GTR correction with rate heterogeneity. Under that model, parsimony dramatically outperformed minimum evolution over much of the parameter space but did worse in the cases of long-branch attraction (see Figure 3.24). If one is relying on distance methods as the sole basis of phylogenetic reconstruction, the four-taxon simulations indicate that simpler models are better. On the other hand if distance methods are being used along with parsimony (as a contrasting method to detect long branch problems that might be misleading parsimony), the simpler models of sequence evolution may lead minimum evolution to fail in the same ways that parsimony does.

P-distances on the sixteen-taxon tree

Use of the most simple distance estimate, the observed percentage of sites which differ between two sequences, dramatically improved the performance of neighbor joining on the sixteen-taxon tree. In fact neighbor joining using simple p distances performed better than either parsimony or maximum likelihood (based on the total number of branches missed). With the simple distance correction, neighbor joining outperformed all of the other methods for all of the branches on the tree except the recognition of the older four taxa groups and the deepest split in the tree (maximum likelihood still performed best for these branches). Steele and Hendy (2000) have shown that, on ultrametric trees, there are good theoretical reasons to expect that p-distances are preferable to complex distance corrections. Such strong analytical reassurances are lacking when ultrametricity cannot be assumed. It should also be noted that the performance of maximum likelihood would almost certainly have dramatically improved had the assumption of a molecular clock been enforced during the inference of these simulated data sets.

Simple Distance Corrections on the Long Angiosperm Tree

The effect of using less complex models was not examined on the short version of the 228-taxon tree because neighbor joining was already doing quite well on this tree. On the long tree neighbor joining with GTR+Γ+I corrected distances was only achieving 70% accuracy even with 16 simulated data sets concatenated end to end. On this tree I tried 16 other types of distance corrections including JC and K2P (with no rate heterogeneity, just invariant sites, just gamma rate heterogeneity and both types of rate heterogeneity), HKY (without rate heterogeneity, with invariant sites only and with both gamma rates and invariant sites), p-distances, and LogDet (without rate heterogeneity and with invariant sites). No distance corrections did as well as the original GTR+Γ+I model for any sequence lengths above 1128 bases. For the shortest simulated sequences there were slight improvements. JC with both types of rate heterogeneity recovered 62% of the tree, and JC with just gamma rates recovered 61% of the tree. These were the only two methods to do better than the original the GTR+Γ+I model (which recovered 60% of the tree). All other corrections performed worse. Neighbor joining using p-distances was a strikingly poor estimator of phylogeny; on average, the use of p-distances would have resulted in the loss of between 15 and 30 branches compared to reconstructions based on distance analyses using the more complex model of evolution.

Conclusions on the Use Simple Distance Corrections

The wide range of effects of using simple distance corrections these three simulations studies underscores the need for clear criteria for determining what model to use. This study offers another example for which simple corrections perform well on ultrametric trees. When a molecular clock cannot be assumed, the situation is less clear. The general conclusions about distance methods in relation to other phylogenetic approaches are not an artifact of the GTR+Γ+I corrections employed being inappropriate. The general guidelines provided by Nei and Kumar (2000) do seem to capture many of the important factors that affect the performance of the distance corrections, but they are too vague to serve as a final answer to the problem.

Implications for Model Improvement from the HB Simulation

In addition to giving information about the overall performance of methods, the HB simulations have allowed me to examine the deficiencies of current models. A benefit of the likelihood approach is that new models can be proposed and then objectively evaluated. Below I will discuss two simple modifications to the GTR model, that were inspired by examining why maximum likelihood failed to reconstruct the extreme Felsenstein zone trees.

Deficiencies of GTR model

The GTR model with rate heterogeneity, probably the most widely used model for maximum likelihood inference, was inconsistent with respect to the tree topology estimation under some conditions. When two non-sister terminal branches had a length of 1.0 and the other three branches had a length of 0.05, the “long-branch attraction tree” was inferred. The branch lengths of the true tree, when estimated from infinite data by the GTR+Γ+I model were quite different from the true branches. The long terminal branches are inferred to have a length of 0.51 instead of 1.0, the shorter terminal branches are overestimated (length 0.063 instead of 0.05), and the internal branch is estimated to be 0.011 (as opposed to 0.05). This tree is a fair approximation of the distance between the short branch taxa (pairwise divergence of 0.137 instead of 0.15), but all of the divergences involving the taxa at the end of long branches are dramatically underestimated. The result of this is the expected frequency of homoplasy uniting the long branch taxa is underestimated.

Properties of the frequency of data patterns, also known as the spectrum, have been explored by Hendy and Penny(1993) and others. For example the Hadamard conjugation uses the observed data to produce a corrected spectrum which takes multiple hits into account. By examining the spectrum of the GTR+Γ+I model when it is attempting to explain data generated under the HB model one can determine what tree and parameter values would be inferred by the GTR+Γ+I model. I have compared the spectra produced by the HB and the GTR+Γ+I models in a heuristic context, as a tool in model development. Figures 7.1, 7.2 and 7.3 show the technique that I have used to visualize the model spectra.

When trying to match data generated by the HB model on the Felsenstein zone tree described above, GTR produces a spectrum with an excess of sites in which all four taxa have A and sites in which all taxa are C, but an insufficient number of constant sites with G or T. Taken alone this would imply that the frequency of G and T in the GTR model should be increased and the frequencies of C and A should be lowered. Clearly the frequencies of other patterns would also be affected by this change of parameters. In particular decreasing the frequency of C in the model would not only decrease the frequency of constant sites with C, it would also decrease the number of patterns in which three taxa have C and the other taxon has a different base. The spectra reveal that while GTR is producing too many sites in which all bases have C, it is not predicting enough sites in which three taxa have a C and the other one has a T or G. The maximum likelihood estimate of the base frequency is a compromise between opposing data partitions (some of which “want” a higher frequency of C and others which would “prefer” a lower value for the frequency of C).

The frequencies of the constant patterns are the largest deviations between the spectra of the two models. This reflects the fact that there are different base frequencies for variable bases compared to constant bases in the HB model. The mutational forces in the HB model fit to cytochrome *b* produce extremely skewed base frequencies (53.7% A, 26.4% C, and 3.9% G). These forces dominate the third base positions where selection is weakest (in fact codon selection is ignored in the model, so selection is often absent in the third base positions). This means that there is a large group of variable sites with base frequencies determined entirely by the mutational forces. At first and second base positions, however, any base can be preserved through the action of selection. This results in a large number of constant sites with base frequencies unrelated to the frequencies in the quickly evolving sites. This phenomenom is not simply an artifact of the HB model. In the real cytochrome b sequences, the average of the empirical base frequencies of first and second base postitions are 24.4% A, 25.1% C, 17.7% G, and 32.9%T; at the third base position the frequencies are radically different (40.4% A, 34.9% C, 3.2% G and 21.4% T).

A Fast Heterogeneous Model of Sequence Evolution

A modification to the GTR model might allow it to effectively address this complication of sequence evolution – simply add a set of base frequencies for the constant sites which are independent of the frequencies at the variable sites. This simple alteration to GTR+I might effectively deal with an aspect of sequence evolution that is probably common in regions where there is heterogeneity in the force of selection. Such a model (GTR+I+IF, for Invariant Frequencies) would only require three additional independent parameters. Furthermore, the model would not be much slower to implement than currently applied models. The inference of three new parameter (which might strongly interact with the branch length and rate heterogeneity parameters), would slow down maximization of the likelihood, but the calculation of the likelihood for a given set of parameter values would take no longer than current models. The GTR+I model is nested within a GTR+I IF model, so model selection could be done with either the likelihood-ratio test statistic (which requires nesting of models), or the Akaike Information Criterion (which does not).

Distance-based methods currently allow the user to remove a proportion of invariant sites based on either the empirical frequencies of the bases or the frequencies of the bases in constant sites only. For the HB simulations, removing constant sites based on their frequency in constant sites produced significantly better trees. This implies that a generalization of the model to likelihood methods might result in an appreciable difference in performance. It may seem paradoxical that adding parameters that only affect the expectation of the number of constants sites of each base could have a profound effect on phylogenetic inference. After all, these sites do not contain signal supporting one phylogeny over the other. Nevertheless, correctly estimating the number of sites which are invariant during evolution, can have profound effects on estimates of branch lengths, and therefore the total amount of homoplasy.

An Approximation of Heterogeneous Selection Pressures

Comparison of the spectra of the GTR+Γ+I model to that HB model reveals other conflicts in the expectated frequency of data patterns. These conflicts might not be adequately addressed by allowing different base frequencies for a class of Invariant sties versus the class of sites which are free to evolve.

When summing the frequencies of all patterns in which the two taxa with short branches differ from each other, there is very little discrepancy between the frequency predicted by the HB model and the frequency predicted by the GTR model. However, the GTR model underestimates the frequency of patterns in which a taxon with a short branch has a base which differs from that found in one of the taxa with a long branch. This would imply that the long branches leading to the divergent taxa are not long enough (which is true). The reason the maximum likelihood estimates of the branch lengths are not longer in the GTR model is revealed when one considers patterns in which the two long branch taxa differ from each other. The frequency of this type of data is much lower in the HB model spectrum than is expected by the GTR model. This is not due to constant sites (the total frequency of constant sites is approximately correct). Particularly striking are the frequencies of the twelve data patterns which are parsimony informative characters supporting the long-branch attraction topology. The expected frequency of all of the characters is dramatically underestimated by the GTR model.

A plausible explanation for the huge excess of misleading homoplasy is the fact that many nucleotide sites are effectively two or three state characters in the HB model. For example, although over the entire data set the frequency of A is high, there are some variable sites which never mutate to A. In one codon almost all taxa code for asparagine (AAY), with a minority having aspartate (GAY); the third base position has a parsimony length of 199 steps on the tree used from inference but no taxon has an A or G at this position. This character would be inferred to have a high rate of evolution by a GTR model, but the amount of convergence would be dramatically understimated. Examination of the spectra indicates that these types of data patterns account for a great deal of the failure of the GTR model to match the data generated by the HB model, so ways to incorporate these patterns should be a high priority.

Halpern and Bruno's approach to this problem was the drastically heterogeneous model implemented in this dissertation, but for the purposes of phylogenetic inference simpler models which only approximate the effects of heterogeneity may be more feasible. An intriguing possibility is to build a family of matrices which are only slightly more parameter rich than the GTR models but which recognize that the "favored" base can vary from site to site. Using the Halpern and Bruno approach, selection can be separated from mutation so that GTR mutational rate parameters form one part of the model. As opposed to inferring a site specific residue frequency (as in the HB model), three parameters describe the degree of selective preference throughout the sequences. The parameters reflect how skewed the base preferences are at each site. They try to capture whether selection usually favors one dominant base and three rare bases or a more even distribution of base frequencies.

There are 24 possible rank orders for the frequency of the bases at a site because any of the four bases can be the most common, any of the remaining three can be the second most common, and either of the other bases can be the third most common base. Given values for the mutational rate parameters and the equilibrium frequency of the dominant base, the equilibrium frequency of the second most common base, and the equilibrium frequency of the third most common base, 24 models can be generated which encompass all possible rank orderings.

The order of base preferences for each site in the sequence is not known *a priori*, so it is unclear which of the 24 models should be applied to a given site. There are several ways to address this problem. One approach would be to choose whichever of the 24 models results in the highest likelihood for a given site. This effectively treats the model choice as another parameter to be estimated. Such a tact would generate many fewer parameters than the HB model, but it would still be very parameter rich and therefore slow and prone to overfitting the data. A second method would take the likelihood of any site to be the mean likelihood of that site based on each of the models. This is similar to how the likelihood is calculated when gamma-distributed rate heterogeneity is approximated using discrete categories, however the approach is much more justifiable for rate heterogeneity because the discrete categories are created by approximating a distribution whose shape is determined by an inferred parameter. The rate categories are created so that a site has an equal *a priori* probability of belonging to any of the categories. In the case of the rank preferences model, there is little reason to believe that the 24 model categories will be found in equal proportions. The most powerful way to address the problem may be to estimate the proportion of sites in each of the 24 categories and then calculate the likelihood of each site as the weighted mean over all categories.   This would require 23 additional parameters to be estimated. All of the approaches would be at least 24 times slower than a GTR likelihood calculation.

It is hard to imagine a model similar to this rank preferences model performing well on data sets with few taxa. The only way robust estimates of the preferences of different sites for different bases can be assessed is if there is enough data to distinguish between similarity of sequences at a site due to phylogenetic inertia and similarity due to selective constraint. Nevertheless, such models may become powerful tools for deeper level phylogeneties of genes which are well sampled. As the sixteen-taxon simulations presented earlier suggest, old branches in the tree may require methods that do a good job of accounting for long branches. Several obvious modifications of the approach described above might be useful. The distinction between the third most frequent base and the fourth might be unimportant; setting these two frequencies equal to each other leads to a family of 12 models with only two parameters controlling the frequencies of the two most common bases. It is conceivable that in many sites two bases are almost equally common and the other two considerably rarer so that only two preference classes need to be created (leading to six models). Perhaps most important would be the creation of one additional category that is free of selective constraint, to account for the potentially large group of sites that reflect the forces of mutation only.

It is possible that neither the rank preferences model nor the invariant frequency model suggested here will prove helpful on real sequences. However, based on the data patterns which the GTR model is failing to explain, both models seem to be worth pursuing.

A Final Result Relevant to Model Building for Phylogenetic Inference

Molecular systematics is experiencing a burst of new model development. In particular there is a trend toward parameter rich models that try to specifically address a widening range of forces of molecular evolution. One of the exciting aspects of the Halpern Bruno model is its distinction between mutation and fixation, opening up the possibility of building more explicit population genetics assumptions into the models of sequence evolution. This could make tree inference more robust or powerful (although neither of those outcomes is guaranteed by the adoption of more complex models) and will almost certainly make molecular phylogenetics much more relevant to the fields of molecular evolution and population genetics. Models such as the GTR model with gamma-distributed rate heterogeneity fundamentally treat the different aspects of molecular evolution as nuisance parameters. For example rate heterogeneity can interfere with robust phylogenetic estimates. The application of gamma distribution to modelling this process was an enormous advance in systematics, but the use of a gamma distribution is not interesting to an evolutionary biologist studying the reasons for rate variation in sequences. It is difficult to incorporate biological knowledge into such a non-mechanistic model of evolution or get useful biological information out of the results. As the models used by phylogeneticists become more mechanistic, there is an increased interest in the values of parameters (as opposed to a sole focus on the inferred topology).

It is relatively easy to add a term to a model which is designed to estimate the effects of some force of molecular evolution. Unfortunately, if the overall model is a gross oversimplification of the real process of evolution (as phylogenetic models generally are), there is the potential that each of the terms in the model reflects a complex mix of forces that create the observed sequence patterns. A simulation based on the HB model provides an interesting cautionary example of how seemingly unrelated parameters in models of evolution can interact and lead to spurious conclusions.

In their review of amino acid and codon models, Yang *et al.* (1998) note that on a data set of 20 mammalian mitochondrial genome sequences, the REV model fits the data significantly better than the REV0 model. Both models are very general models of amino acid evolution. The difference between the two is that the REVO model disallows changes between amino acids whose codons are not one mutational step away from each other. Yang *et al.* state that the comparison of the two models "constitutes a test of the hypothesis that amino acid (codon) substitutions proceed in a stepwise manner, with each step involving a change at only one codon position." Their rejection of REV0 is intriguing because it indicates an important role for an unusual type of non-independence of the nucleotide substitutions. They do note that "possible factors [leading to interdependence of substitutions] are mutations affecting more than one nucleotide site, compensatory nucleotide substitutions, and selective pressures at the DNA level." This hypothesis test, while not the focuse of their paper, is a good example of a new application of phylogenetic models with surprising implications for how gene sequences evolve.

While the REV and REV0 models are quite general, they do assume that molecular evolution is homogenous across the sequence. The most obvious type of signal that would cause an analysis to reject REV0 in favor REV are the presence of sites in which some species have an amino acid two mutational steps away from the amino acid found in other species and no taxon with any of the intermediate amino acids is observed (despite the fact that, at other sites in the molecule, the intermediate amino acids interchanges with each of the other two amino acids). While mutations affecting multiple sites would be one way to explain this pattern, another explanation is that, at some sites in a protein, amino acids that are not mutationally adjacent are preferred, but, when an intermediate amino acid is found, the protein does not function well. This type of pattern can be produced by the HB model.

To test if the heterogeneous amino acid preferences of the HB model are strong enough to result in REV0 being rejected, I simulated data using the HB model onto the tree used by Yang *et al.* To mimic the amount of data they used, I simulated multiple copies of the cytochrome *b* gene. Despite the fact that amino acid evolution occurs in a stepwise fashion in the version of the HB model that I simulated under (no multiple mutations or other unusual nucleotide-level non-independence is modelled), REV0 was rejected in favor of REV on the simulated data set. It is possible that multiple mutations or a similar form of non-independence is responsible for the rejection of REV0 in Yang *et al.*'s study, but it is interesting that a seemingly very different phenomenon (heterogeneous selection pressures across sites) can be mistaken for evolution not proceeding in a stepwise manner. The example underscores the need for caution when using very simple models to infer the evolutionary processes responsible for the patterns that we see in sequences.

Search	Type of parsimony	Higher Level Constratint	Cutoff for Nodes from the Previous Search	Type of Swap
1a	TiTv	Deep	NA	None
1b	Unord	None	NA	None
1c	TiTv	Deep	NA	None
1d	Unord	Deep	NA	None
2a	TiTv	None	90	SPR
2b	Unord	None	90	SPR
3a	TiTv	Deep	90	SPR
3b	Unord	Deep	90	SPR
4a	TiTv	Deep	90	TBR
4b	Unord	Deep	90	TBR
5a	TiTv	Waddell	90	TBR
5b	Unord	Waddell	90	TBR
6a	TiTv	Waddell	90	TBR
6b	Unord	Waddell	90	TBR
7a	TiTv	Waddell	90	TBR
7b	Unord	Waddell	90	TBR

Table 2.1

Search	Type of parsimony	Higher Level Constratint	Cutoff for Nodes from the Previous Search	Type of Swap
8a	TiTv	Waddell	90	TBR
8b	Unord	Waddell	90	TBR
9a	TiTv	Waddell	80	TBR
9b	Unord	Waddell	80	TBR
10a	TiTv	Waddell	70	TBR
10b	Unord	Waddell	70	TBR
11a	TiTv	Waddell	70	TBR
11b	Unord	Waddell	70	TBR

Table 2.1 (cont.)

Branch	Best Method(s)	Second Group	Third Group	Worst
Young Clades	MP (100%) ML (100%) NJ (100%) ME (98%)			
Young 1-2	MP (98%)	ML (85%)	NJ (59%)	ME (48%)
Young 1-2-3	MP (91%)	ML (62%)	NJ (57%)	ME (49%)
Old Clades	ML (84%)	MP (74%)	NJ (65%) ME (64%)	
Old 1-2	MP (99%) ML (98%)	NJ (87%)	ME (83%)	
Old 1-2-3	ML (90%) MP (87%)	NJ (66%)	ME (61%)	
Deep (Felsenstein)	ML (51%)	MP (38%)	NJ (34%) ME (32%)	
Deep (Farris)	ML (46%)	MP (42%)	NJ (32%) ME (32%)	

Table 4.1

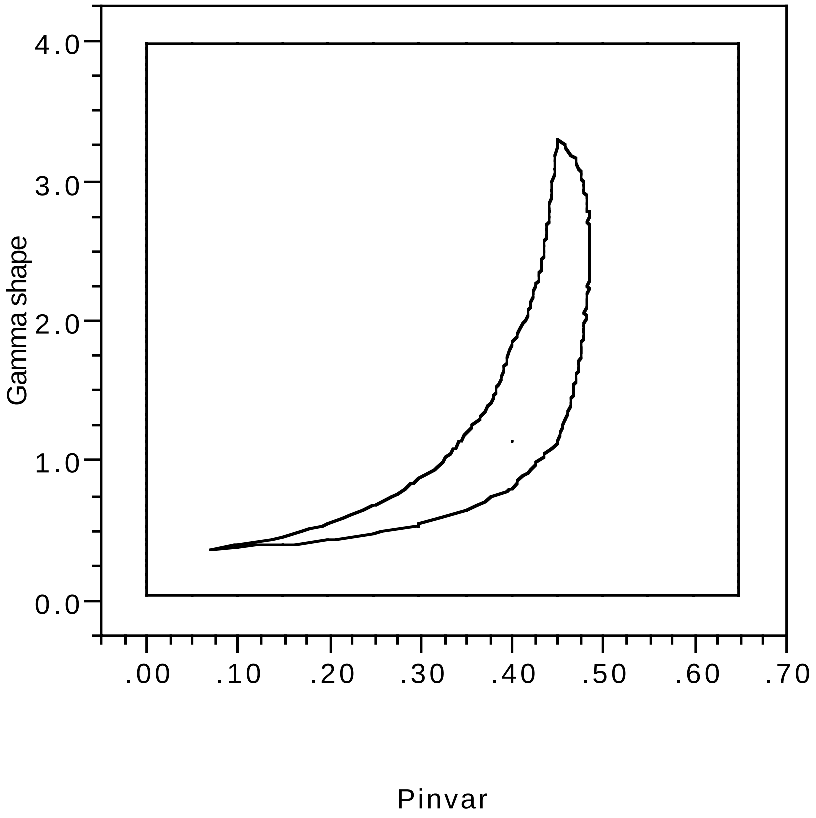


Figure 1.1

A contour plot of the likelihood surface showing the interaction of the estimate of the proportion of Invariant sites (pinvar) and the shape parameter (α) of the gamma distribution of rates across variable sites. The maximum likelihood estimate of the parameters marked by a dot at a pinvar=0.4 and α=1.15. The contour line encompasses all values that would not be rejected using a likelihood-ratio test when compared to the maximum likelihood value. If either pinvar or the α is fixed, the other parameter has only moderately sized confidence intervals. Because of their strong interaction, if neither is known a wide range of values (0.06<pinvar<0.475 and 0.45 < α <3.4) are plausible.

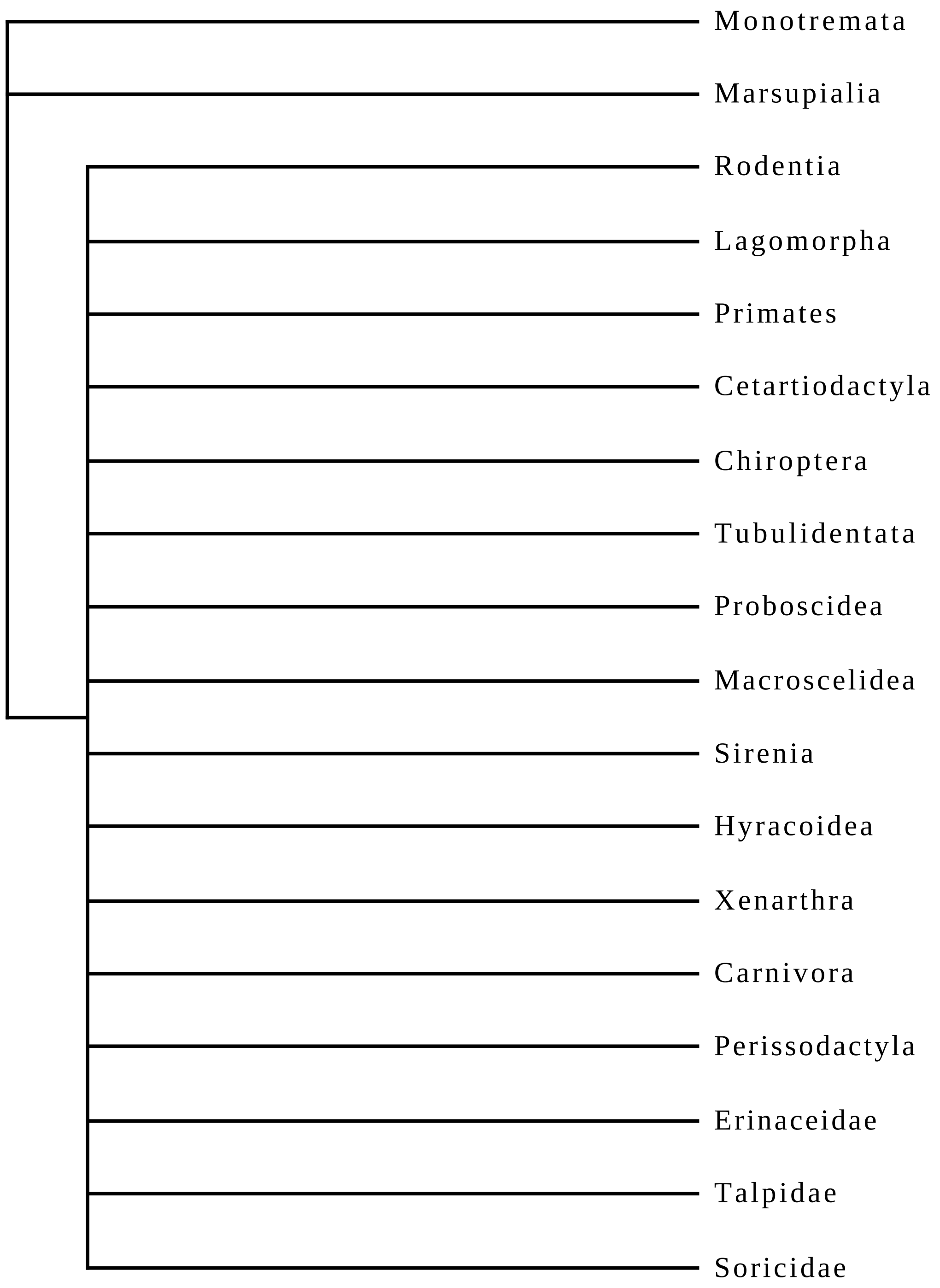

Monotremata
Marsupialia
Rodentia
Lagomorpha
Primates
Cetartiodactyla
Chiroptera
Tubulidentata
Proboscidea
Macroscelidea
Sirenia
Hyracoidea
Xenarthra
Carnivora
Perissodactyla
Erinaceidae
Talpidae
Soricidae

Figure 2.1

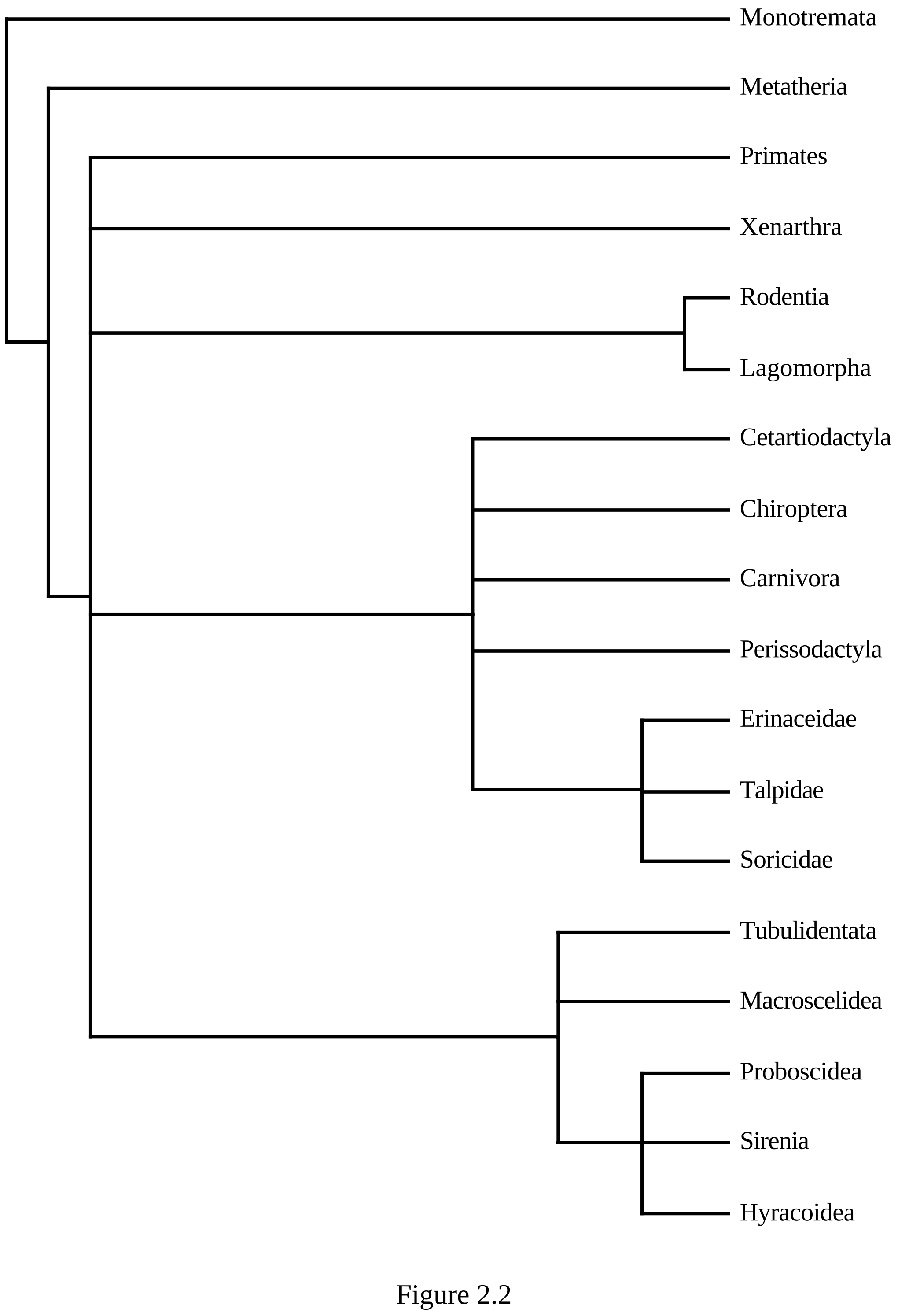

Monotremata
Metatheria
Primates
Xenarthra
Rodentia
Lagomorpha
Cetartiodactyla
Chiroptera
Carnivora
Perissodactyla
Erinaceidae
Talpidae
Soricidae
Tubulidentata
Macroscelidea
Proboscidea
Sirenia
Hyracoidea

Figure 2.2

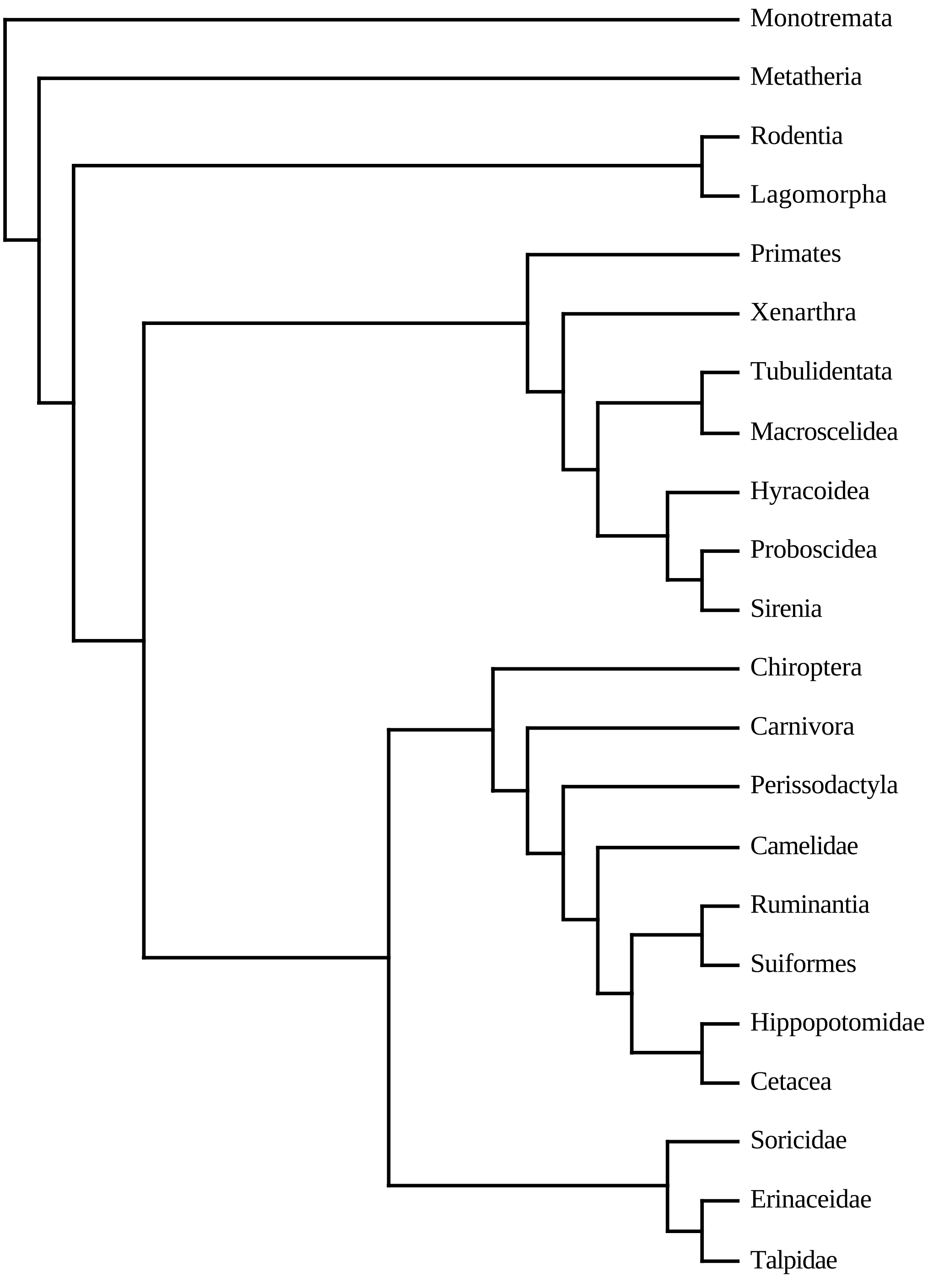

Monotremata
Metatheria
Rodentia
Lagomorpha
Primates
Xenarthra
Tubulidentata
Macroscelidea
Hyracoidea
Proboscidea
Sirenia
Chiroptera
Carnivora
Perissodactyla
Camelidae
Ruminantia
Suiformes
Hippopotomidae
Cetacea
Soricidae
Erinaceidae
Talpidae

Figure 2.3

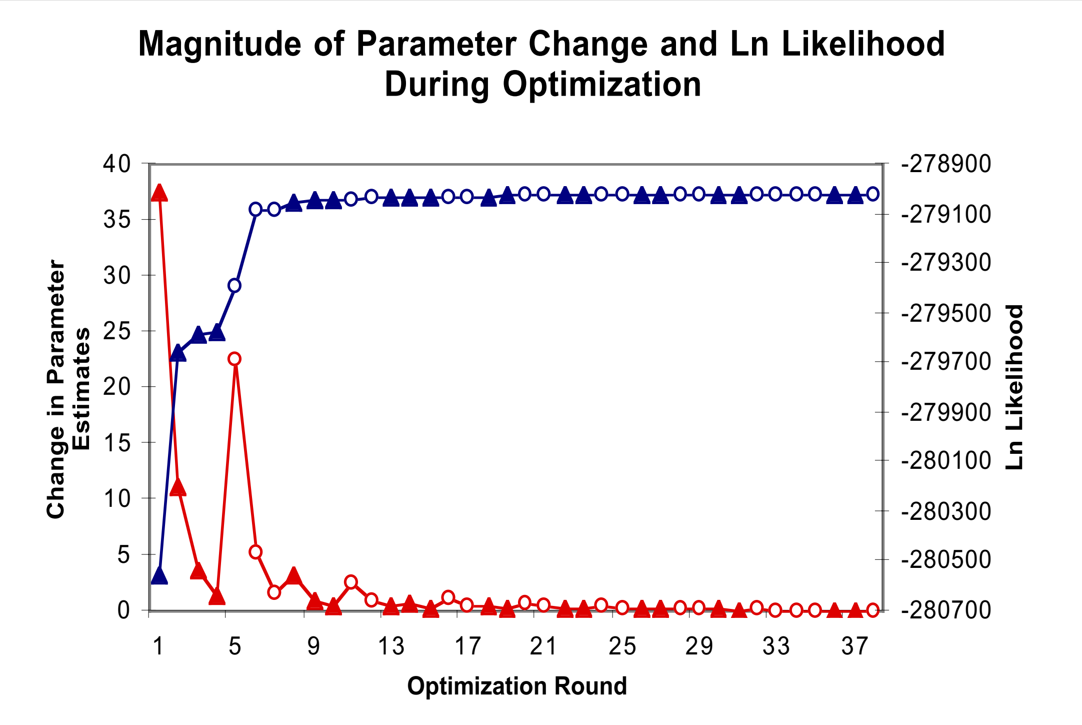


Figure 2.4

A plot of the Ln Likelihood of the HB model fit to the cytochrome b sequences in blue and the change in the parameters in red both as a function of the round of parameter optimizaton for all 38 rounds of optimization. The change in parameters was quantified as the Euclidean distance in parameter space moved in the course of a round. Triangles indicate rounds in which model parameters were changed. The circles indicate rounds of branch length optimization.

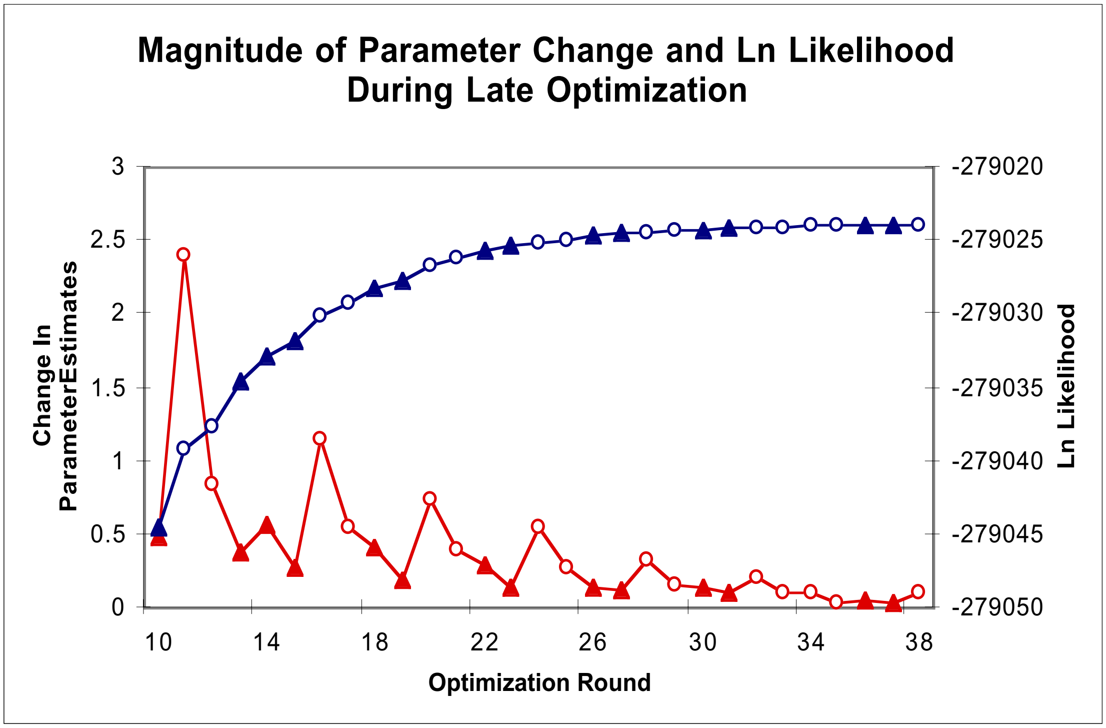


Figure 2.5

A plot of the Ln Likelihood of the HB model fit to the cytochrome b sequences in blue and the change in the parameters in red both as a function of the round of parameter optimizaton for the last 28 rounds of optimization. The change in parameters was quantified as the Euclidean distance in parameter space moved in the course of a round. Triangles indicate rounds in which model parameters were changed. The circles indicate rounds of branch length optimization.

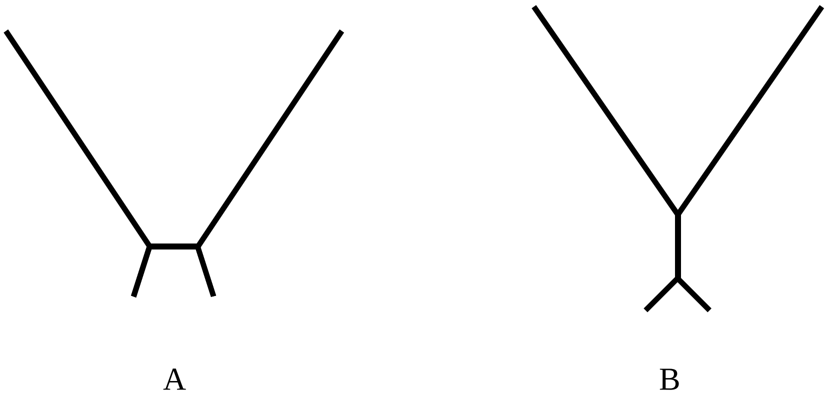


Figure 3.1

Tree A is the Felsenstein zone tree for which many methods are inconsistent. The long branches cause two non sister terminals to be inferred as each other's closest relative. Tree B is the Farris zone tree, which is correctly inferred by parsimony with very little data, but requires more characters to be inferred correctly by distance or likelihood approaches.

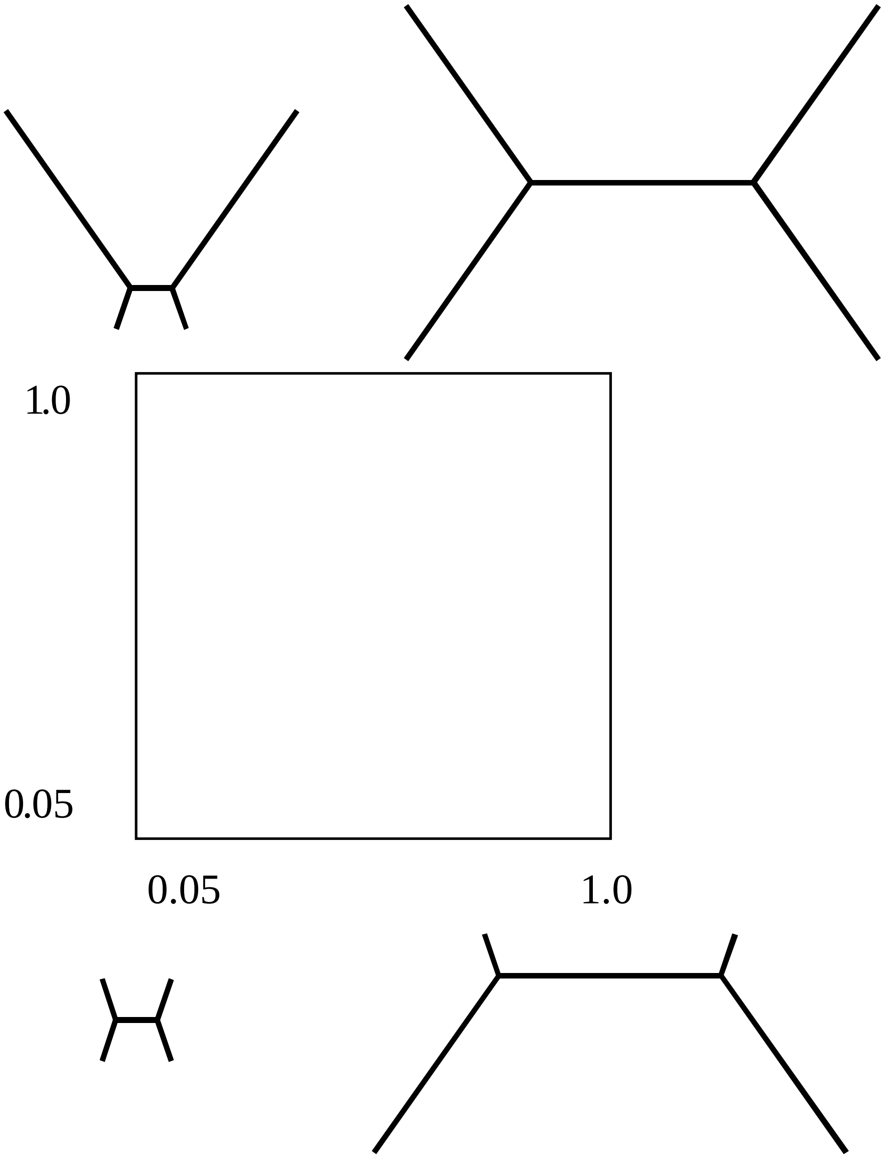


Figure 3.2

The paramter space of the four taxon simulations. The vertical axis is the two-branch length. The horizontal branch is the three-branch length. Approximate tree shapes for each of the corners are shown. Branch lengths between 0.05 and 1.0 expected changes per site were considered.

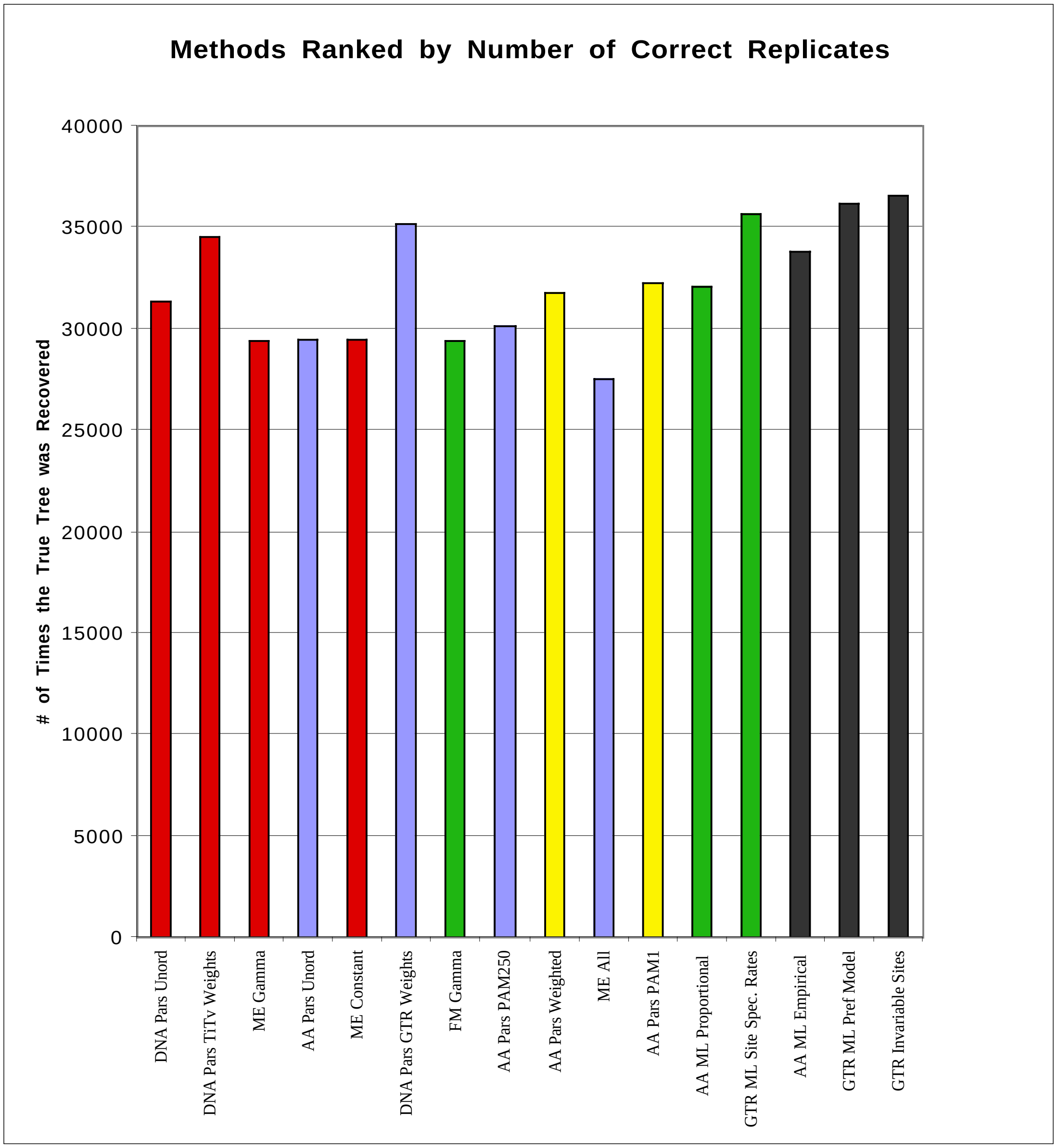


Figure 3.3

A bar graph of the number of times the true tree was recovered by each of the methods over the whole four taxon parameter space. The total number of replicates was 40,000. Distance methods are shown in red, DNA parsimony methods in blue, DNA-based likelihood methods in black, Amino acid parsimony methods in yellow, and amino acid likelihood methods in green.

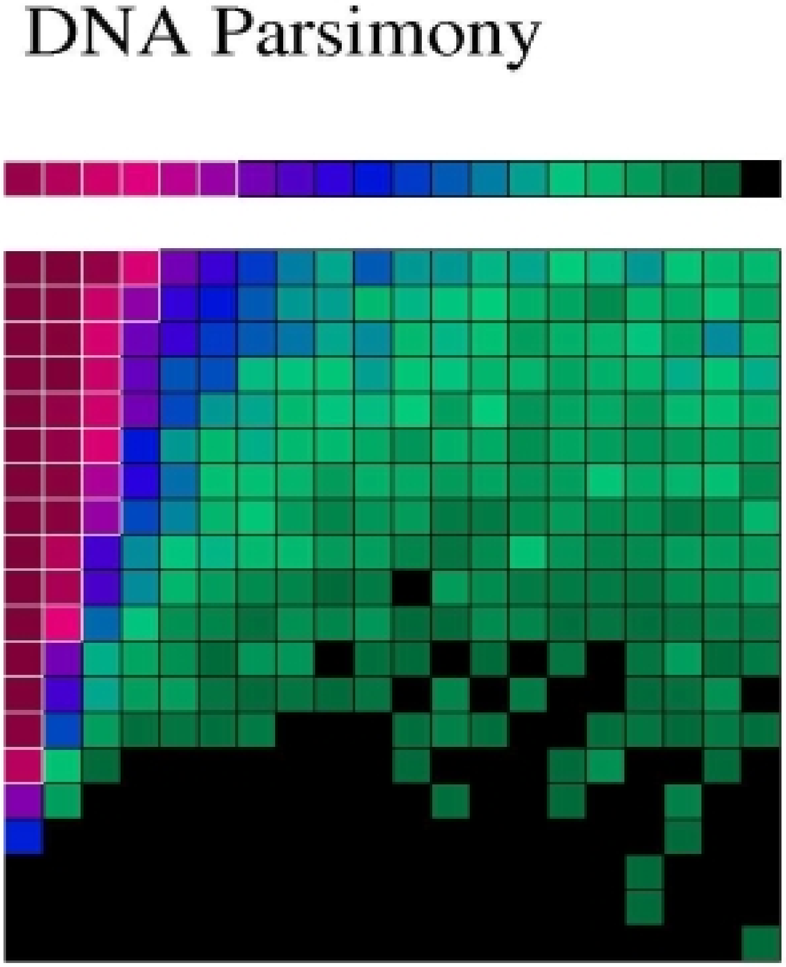


Figure 3.4

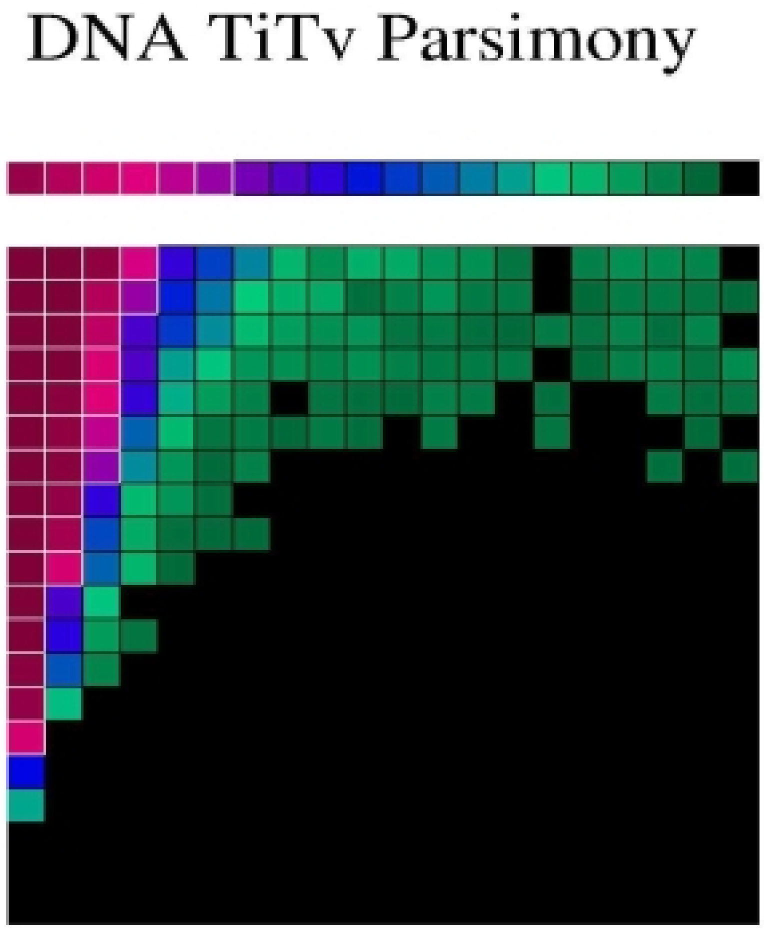


Figure 3.5

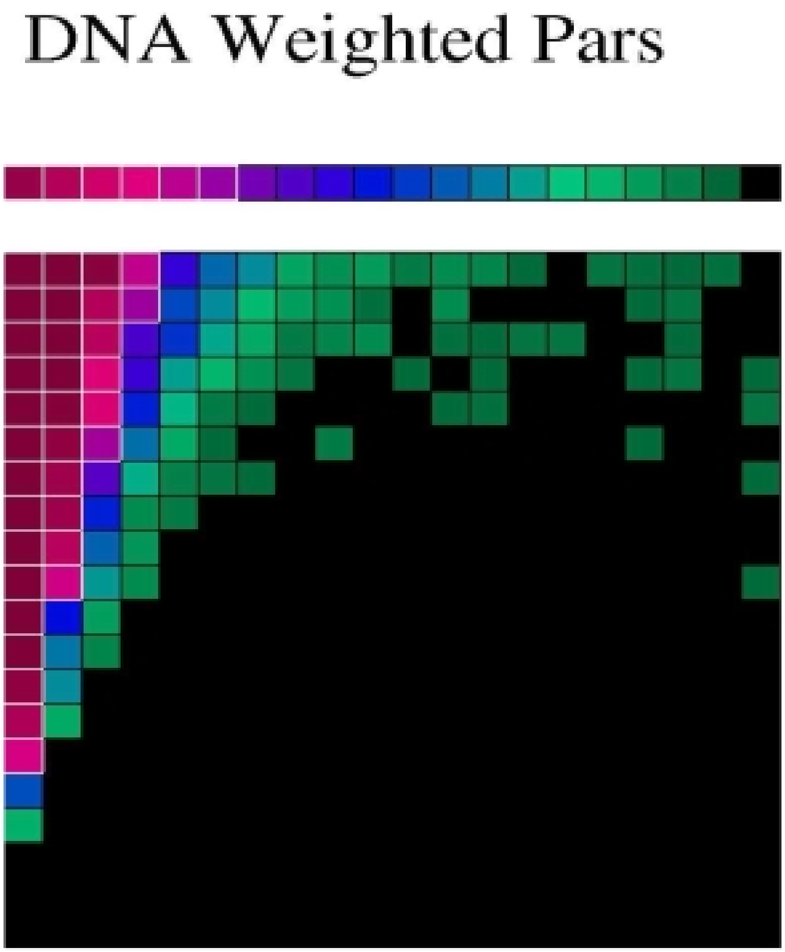


Figure 3.6

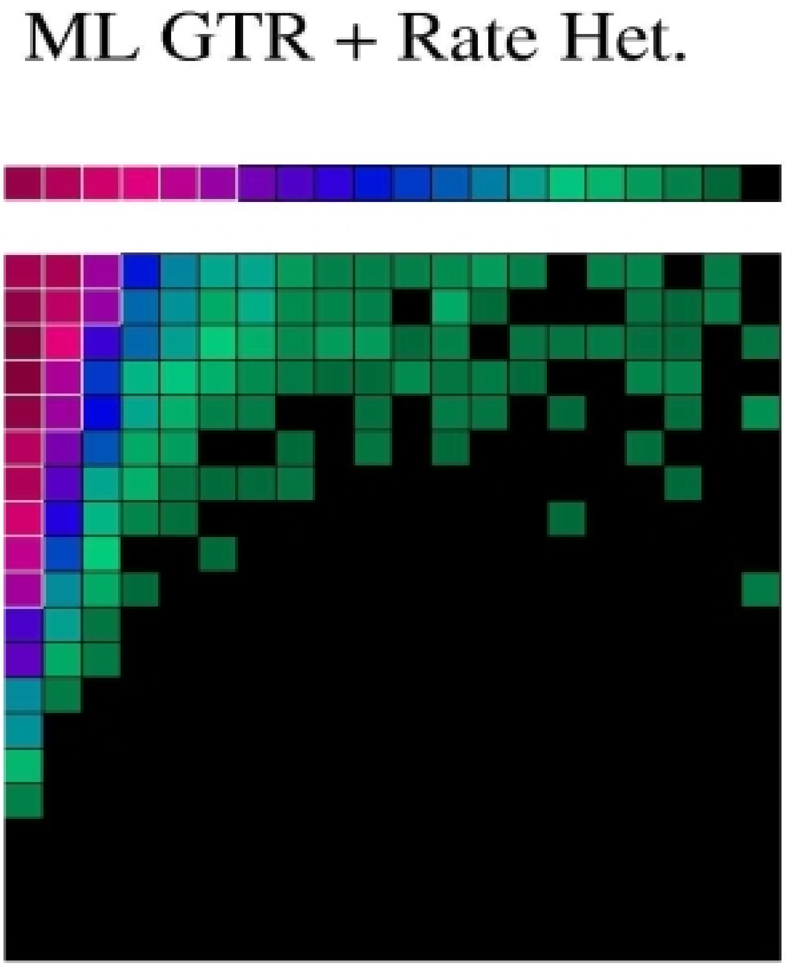


Figure 3.7

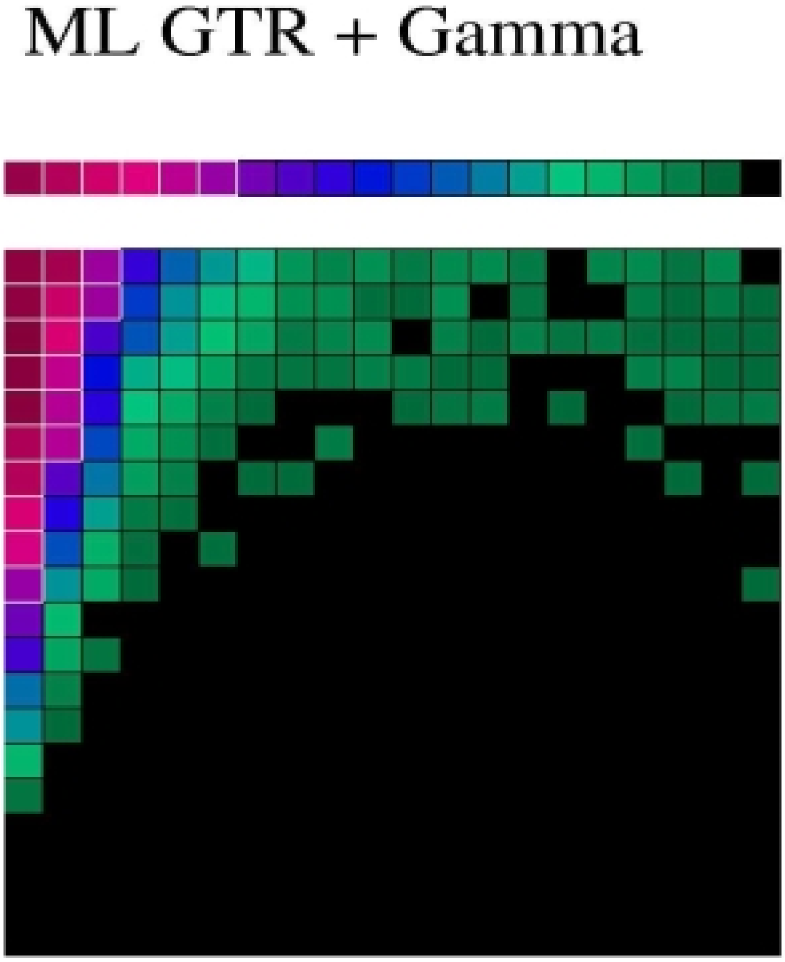


Figure 3.8

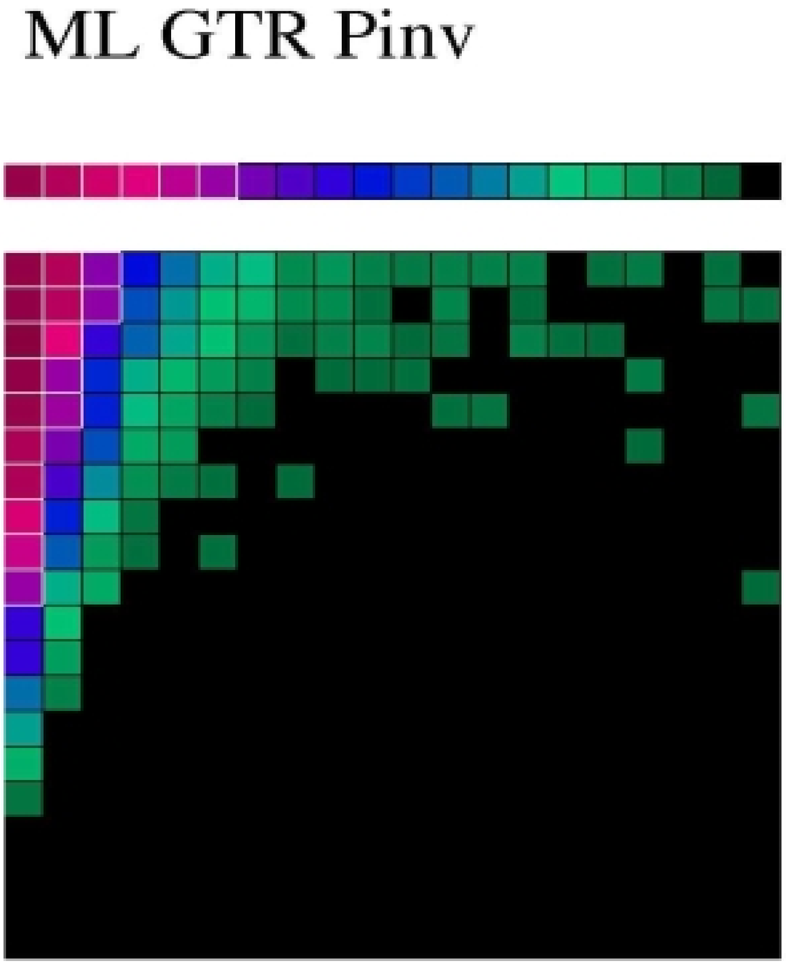


Figure 3.9

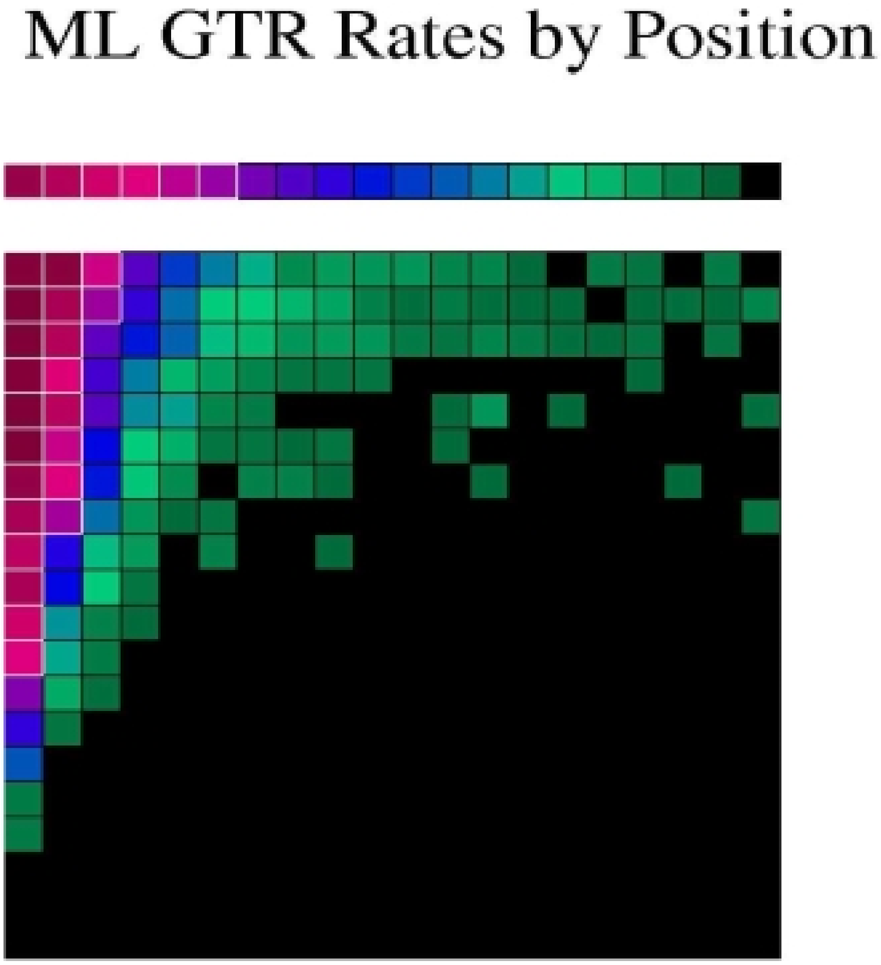


Figure 3.10

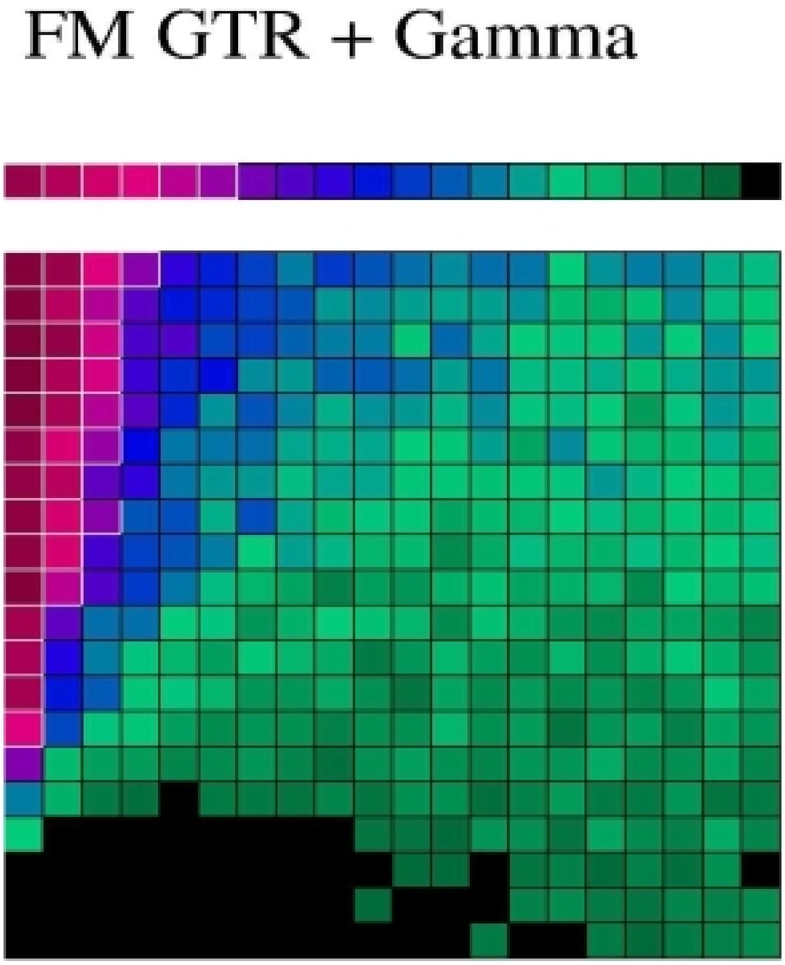


Figure 3.11

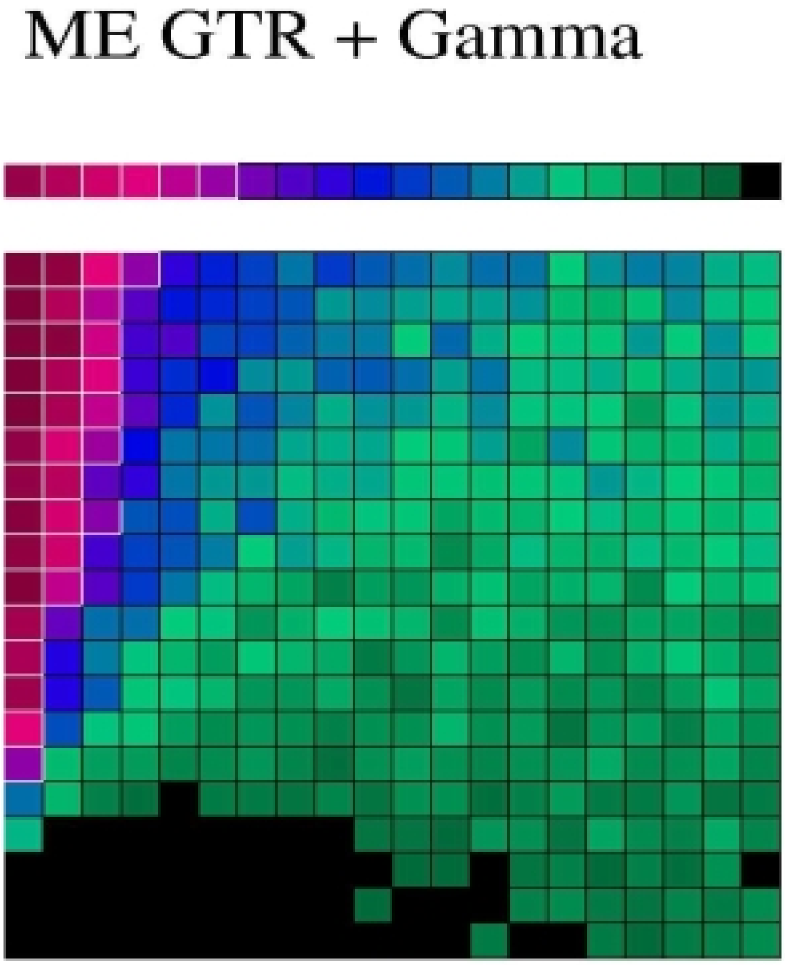


Figure 3.12

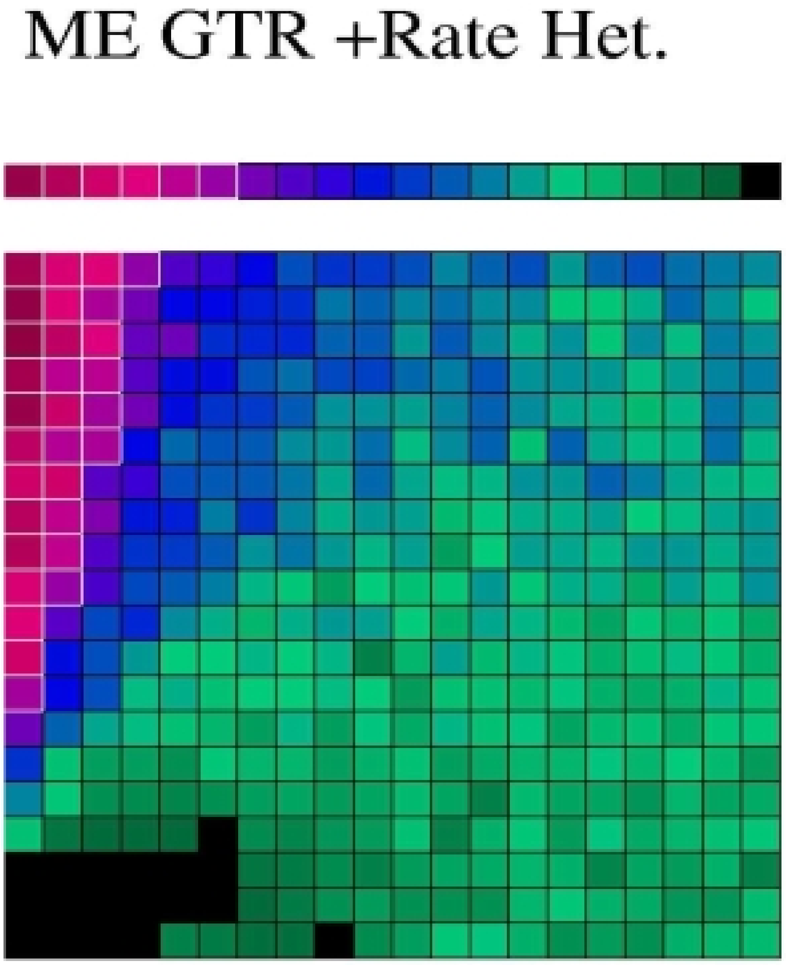


Figure 3.13

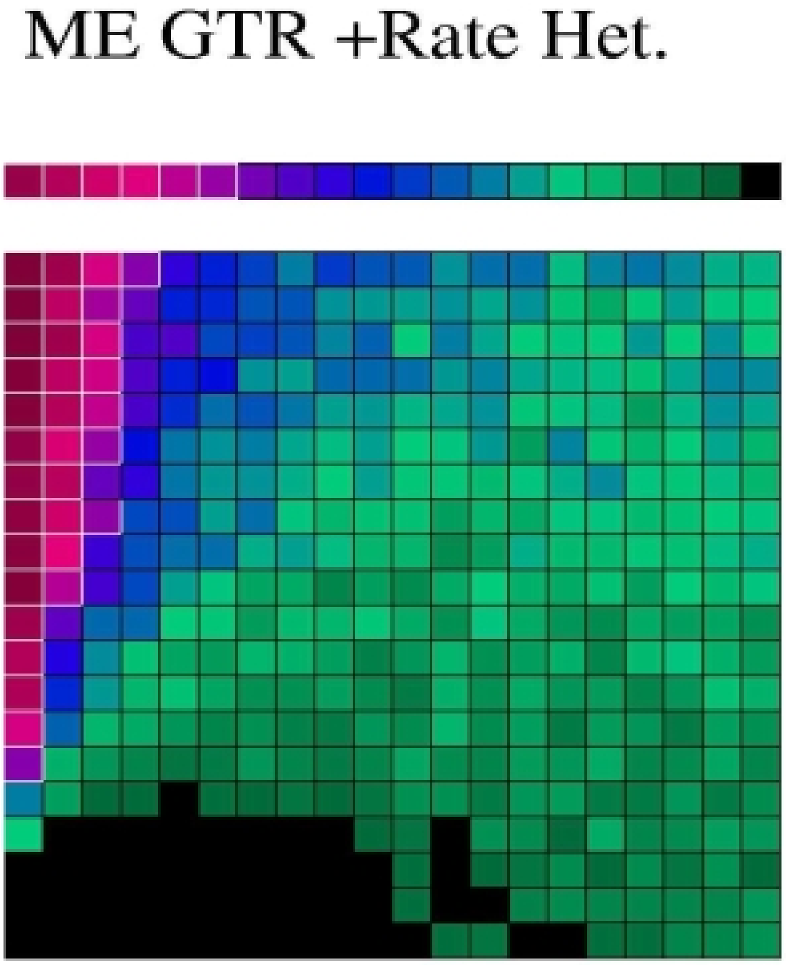
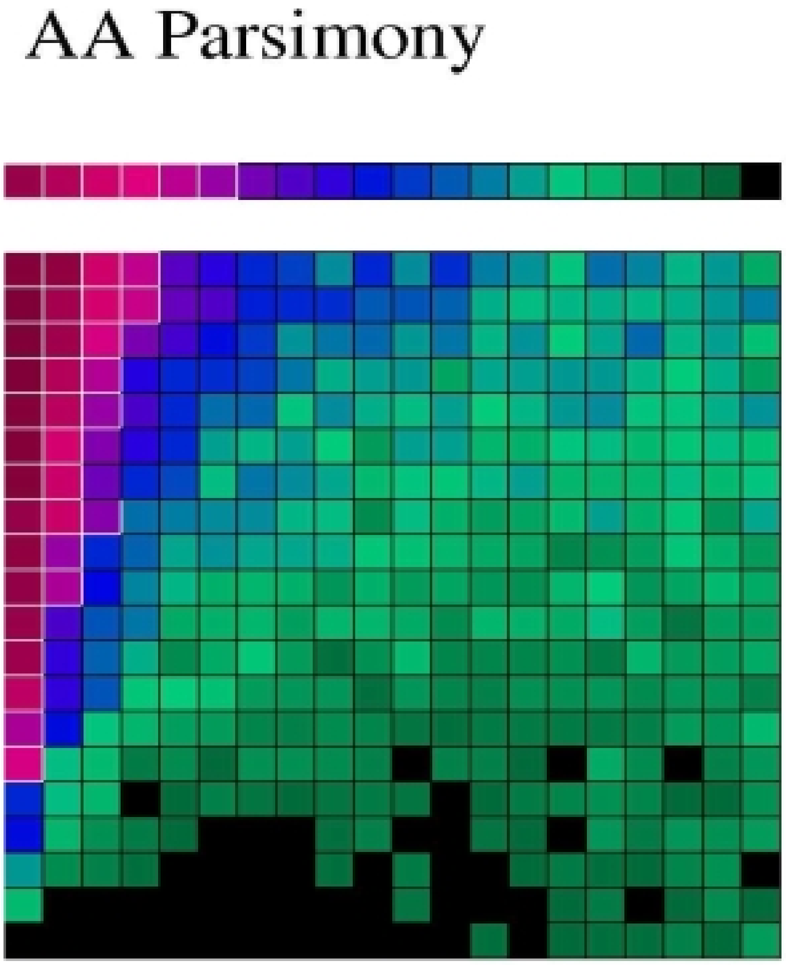


Figure 3.14



AA Parsimony

Figure 3.15

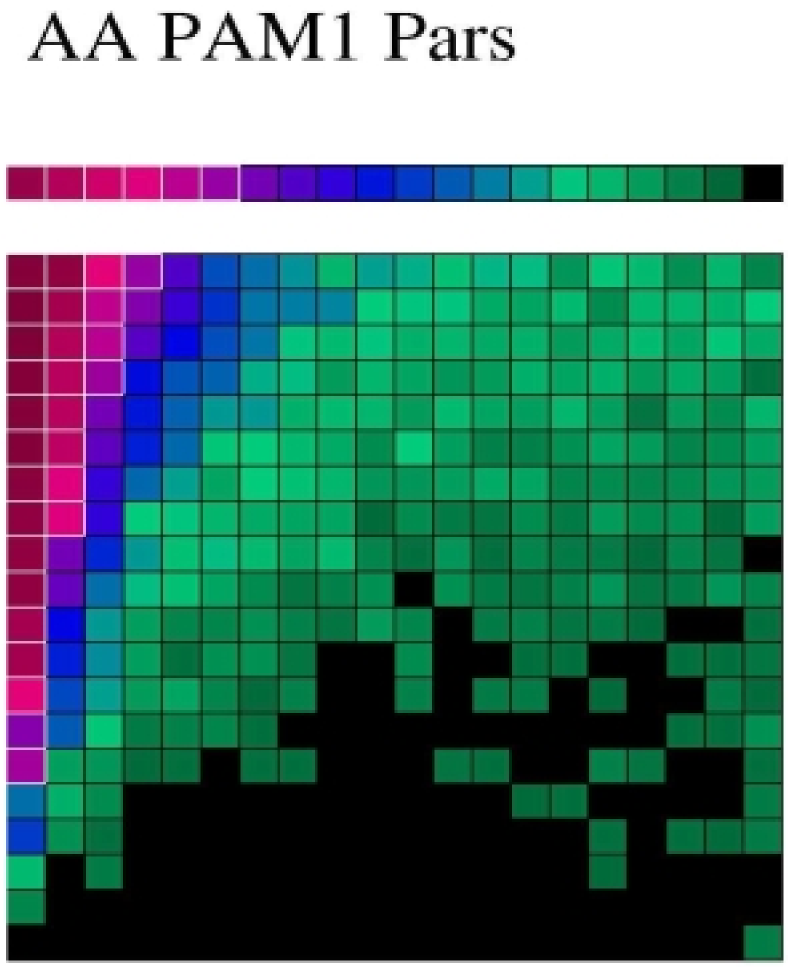


Figure 3.16

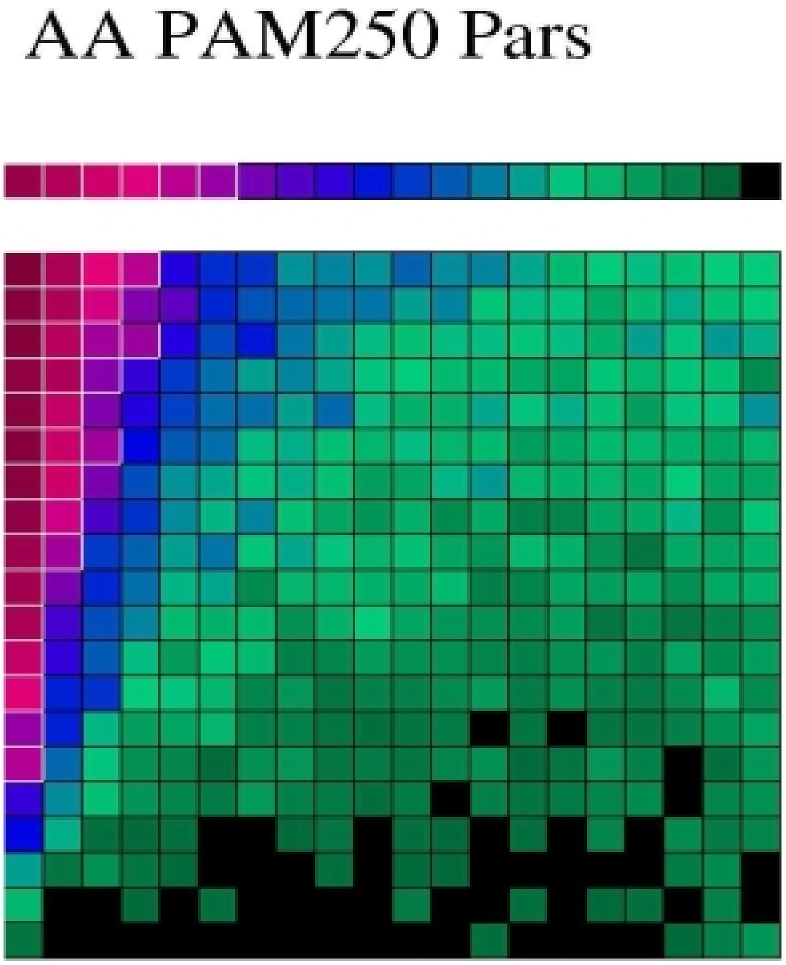


Figure 3.17

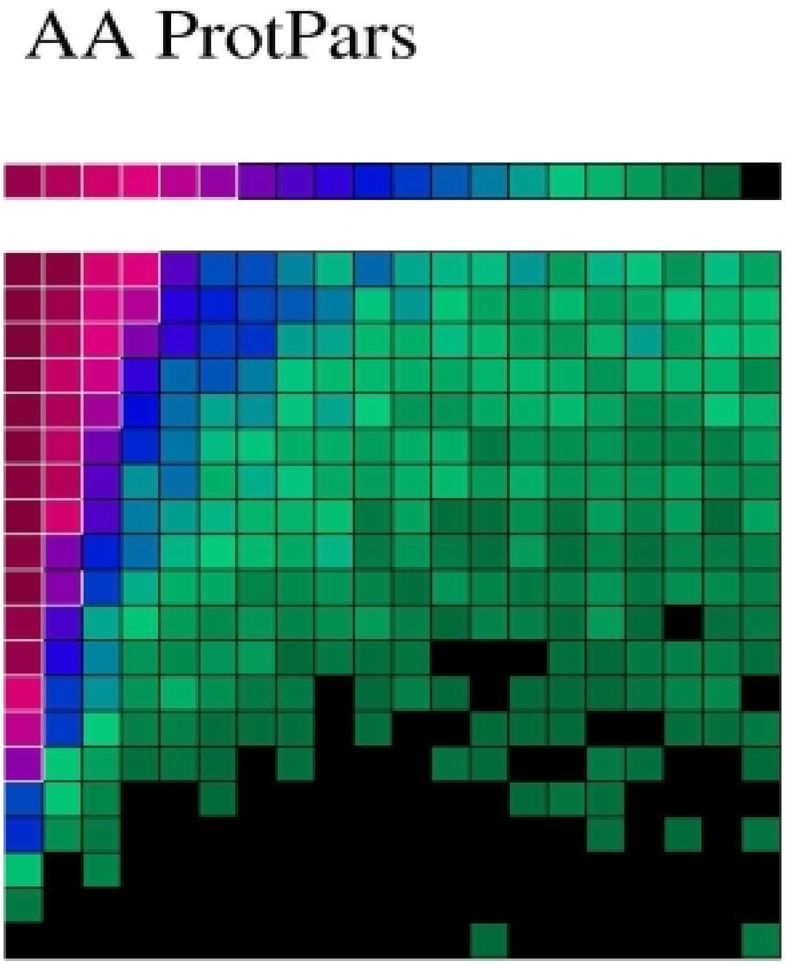


Figure 3.18

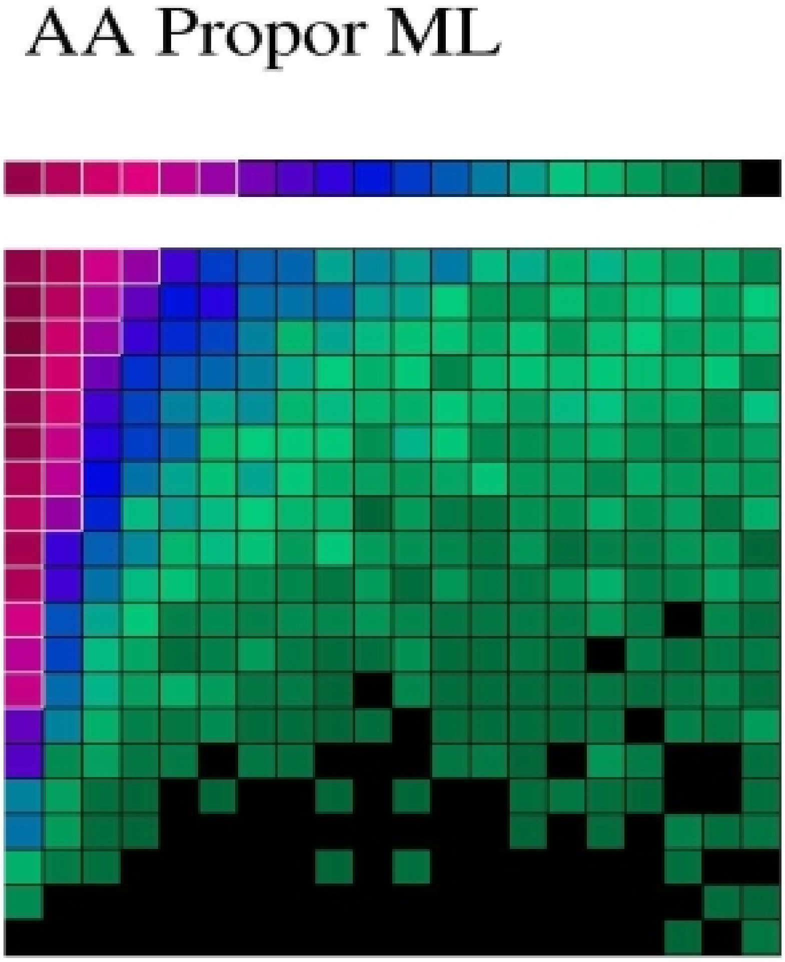


Figure 3.19

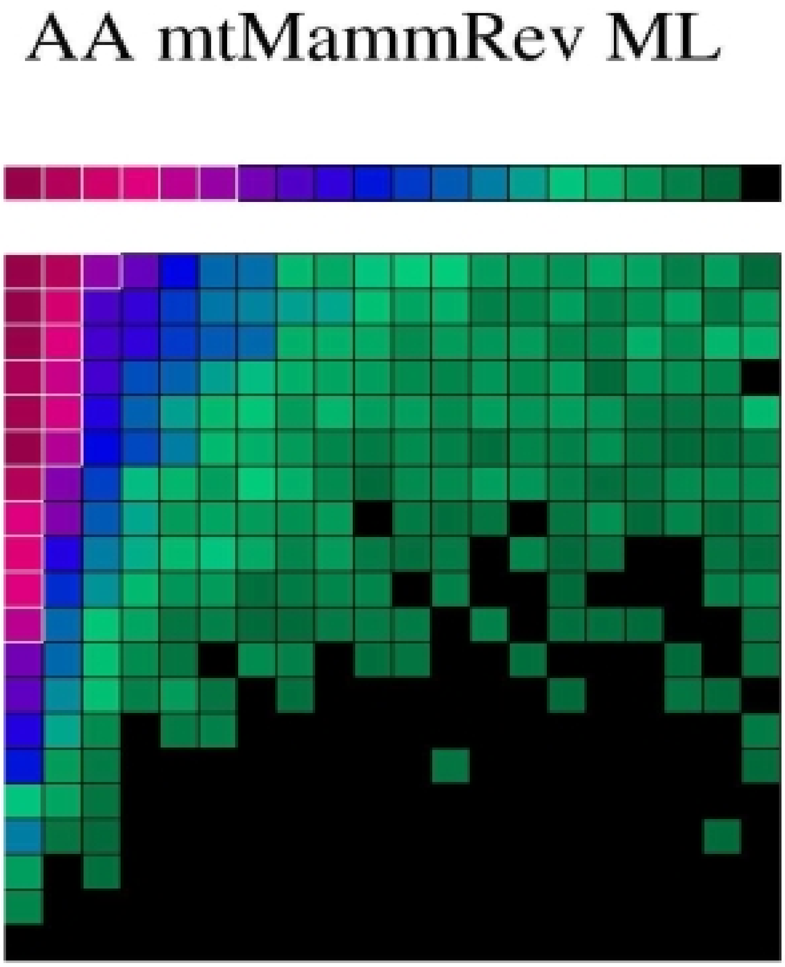


Figure 3.20

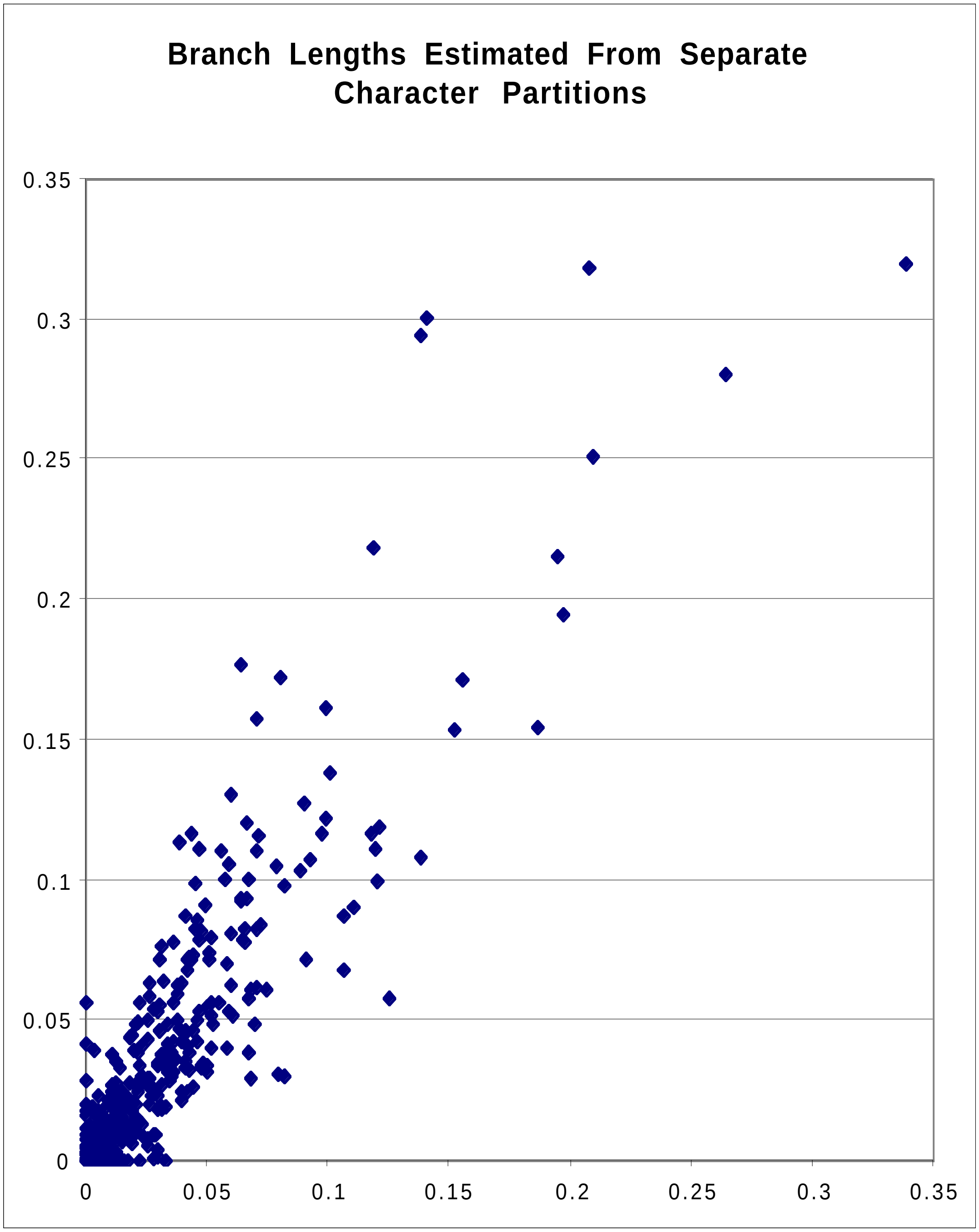

Branch Lengths Estimated From Separate Character Partitions
0.35
0.3
0.25
0.2
0.15
0.1
0.05
0
0
0.05
0.1
0.15
0.2
0.25
0.3
0.35

Figure 3.21

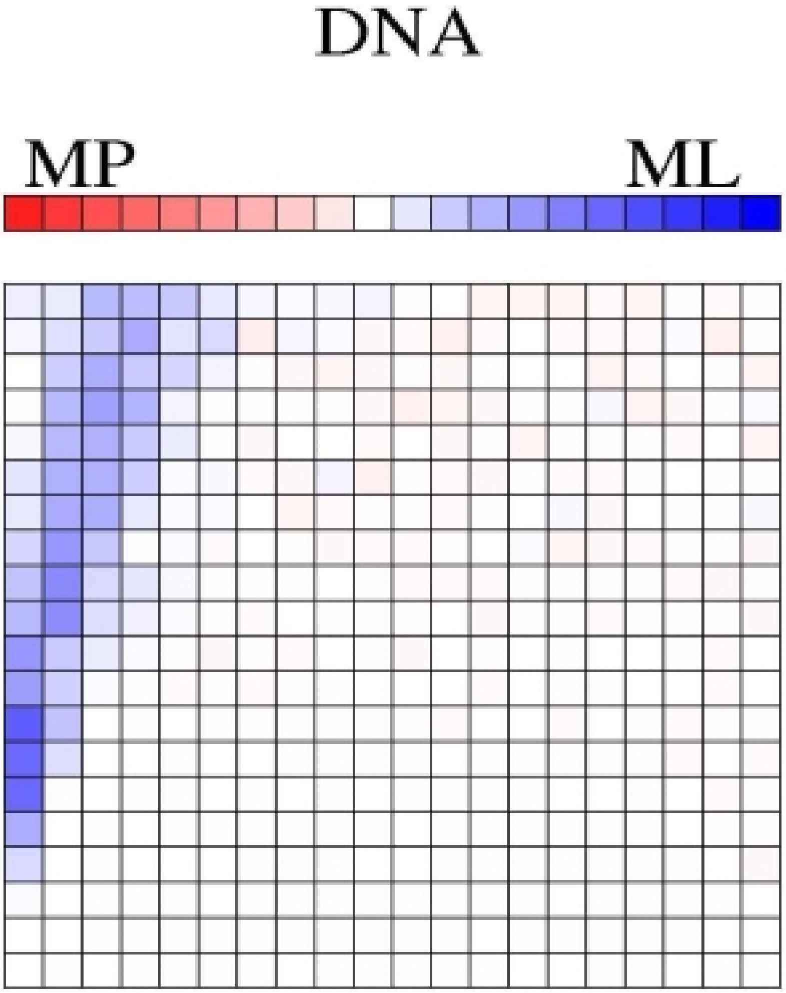


Figure 3.22

A contrast of DNA-based weighted parsimony and maximum likelihood using GTR with preferred rate heterogeneity made by subtracting the number of successes by one method from the number of successes by the other method across the paremeter space shown in Figure 3.2. The maximum value is 100 (shown in deep blue), the minimum value is –100 shown as deep red. White indicates equivalent performance.

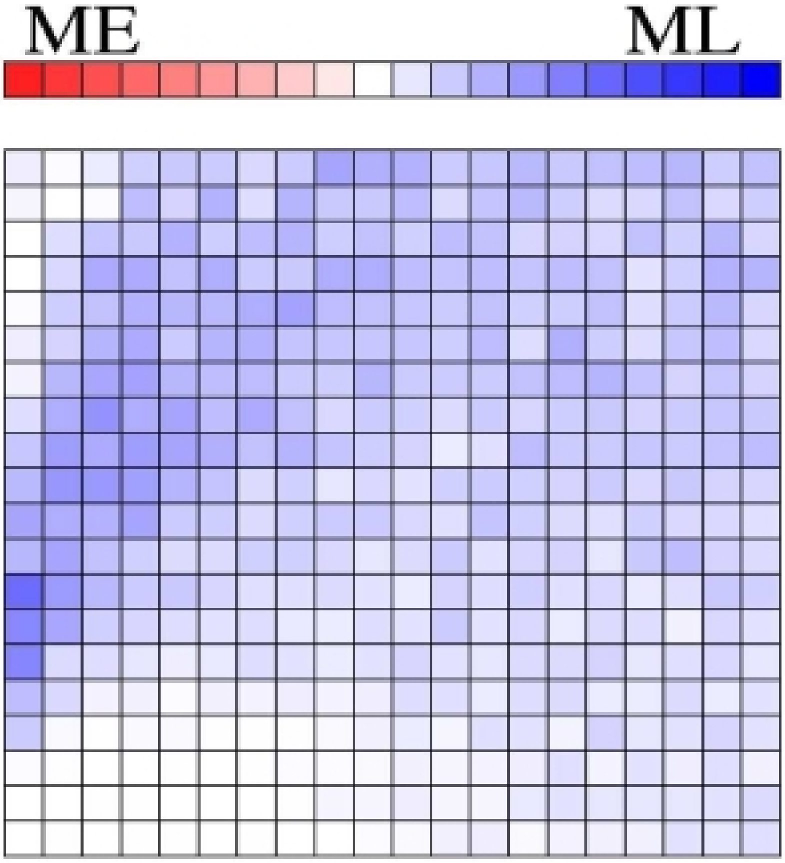


Figure 3.23

A contrast of DNA minimum evolution and maximum likelihood (both using the GTR model with preferred rate heterogeneity) made by subtracting the number of successes by one method from the number of successes by the other method across the paremeter space shown in Figure 3.2. The maximum value is 100 (shown in deep blue), the minimum value is –100 shown as deep red. White indicates equivalent performance.

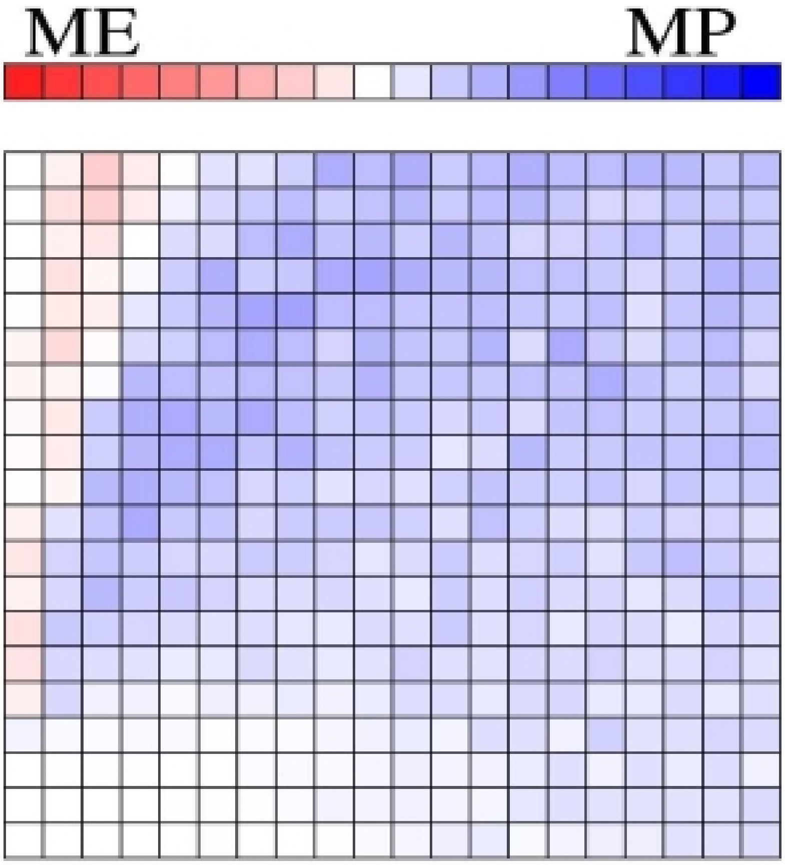


Figure 3.24

A contrast of DNA minimum evolution (using the GTR model with preferred rate heterogeneity) and DNA-based weighted parsimony made by subtracting the number of successes by one method from the number of successes by the other method across the paremeter space shown in Figure 3.2. The maximum value is 100 (shown in deep blue), the minimum value is –100 shown as deep red. White indicates equivalent performance.

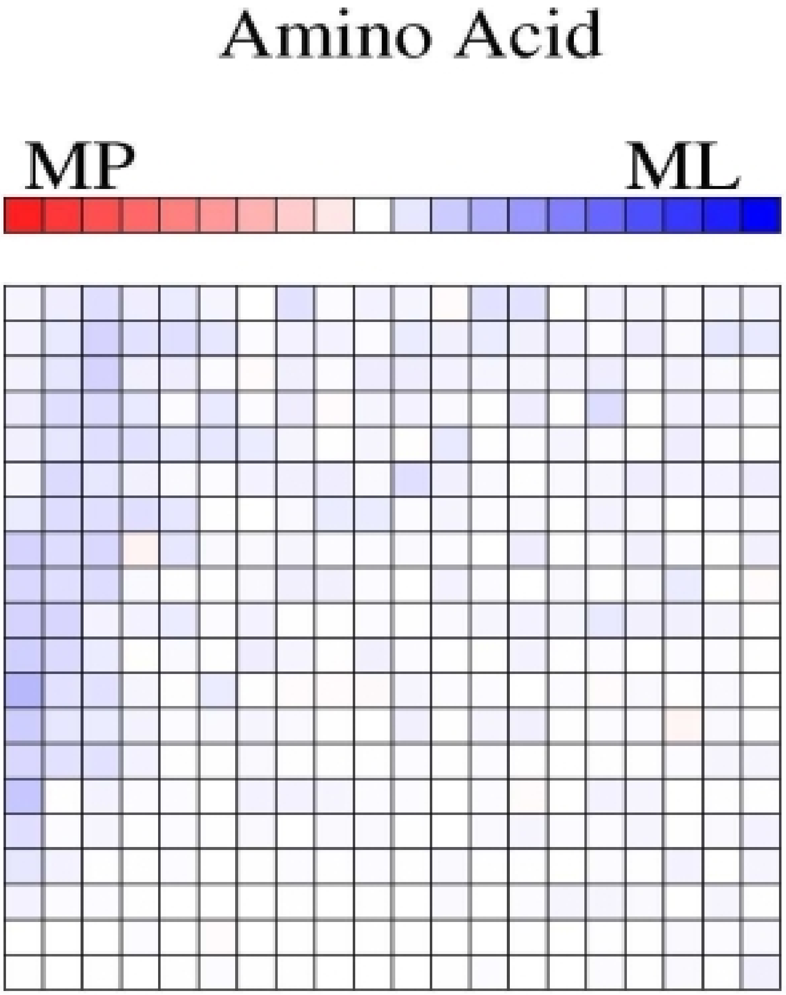


Figure 3.25

A contrast of amino acid parsimony (PAM1 matrix) and amino acid likelihood (using the mtMammREV model) made by subtracting the number of successes by one method from the number of successes by the other method across the paremeter space shown in Figure 3.2. The maximum value is 100 (shown in deep blue), the minimum value is –100 shown as deep red. White indicates equivalent performance.

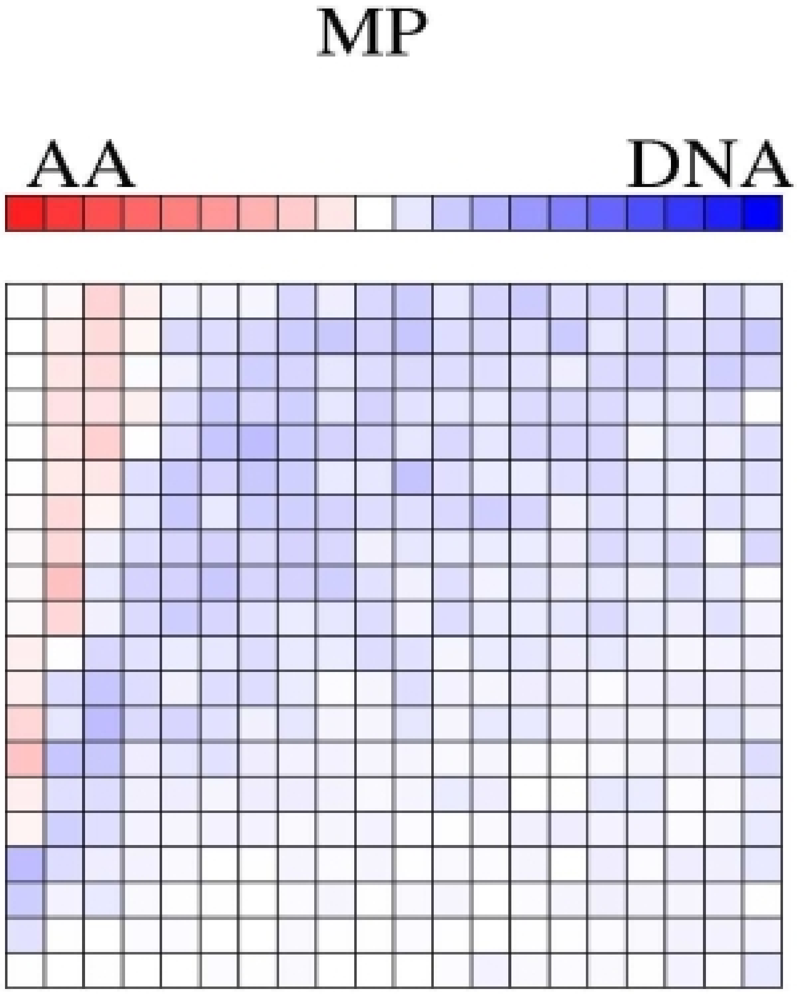


Figure 3.26

A contrast of amino acid weighted parsimony (PAM1 matrix) and DNA weighted parsimony made by subtracting the number of successes by one method from the number of successes by the other method across the paremeter space shown in Figure 3.2. The maximum value is 100 (shown in deep blue), the minimum value is –100 shown as deep red. White indicates equivalent performance.

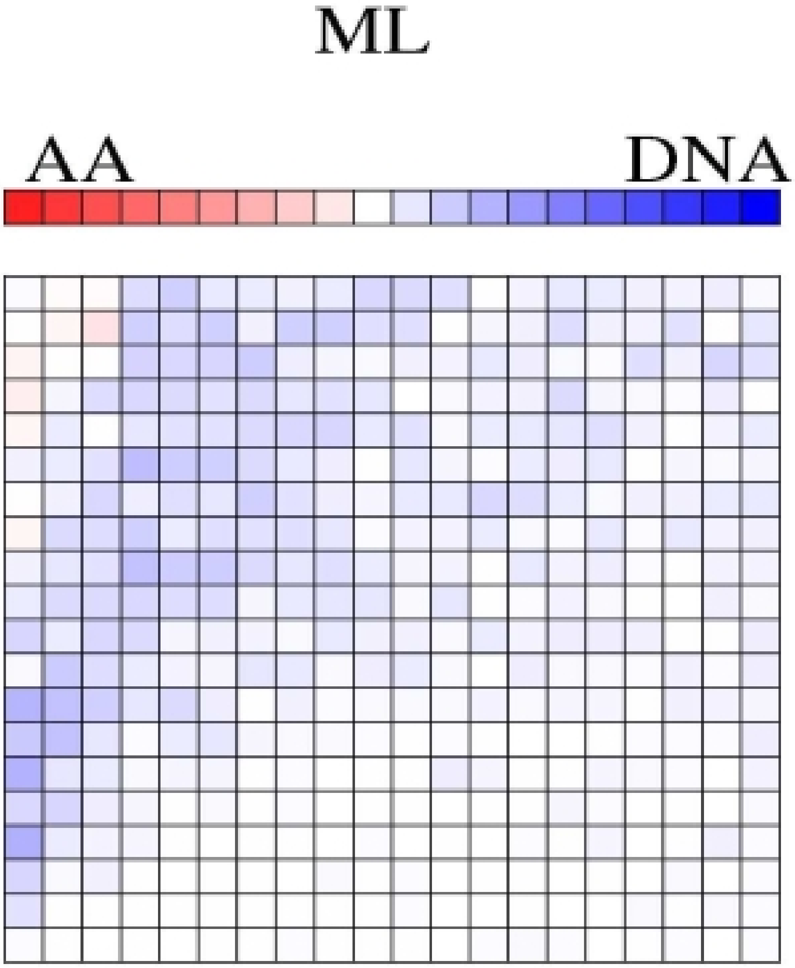


Figure 3.27

A contrast of amino acid likelihood (mtMammREV model) and DNA-based maximum likelihood (GTR model with preferred rate heterogeneity) made by subtracting the number of successes by one method from the number of successes by the other method across the paremeter space shown in Figure 3.2. The maximum value is 100 (shown in deep blue), the minimum value is –100 shown as deep red. White indicates equivalent performance.

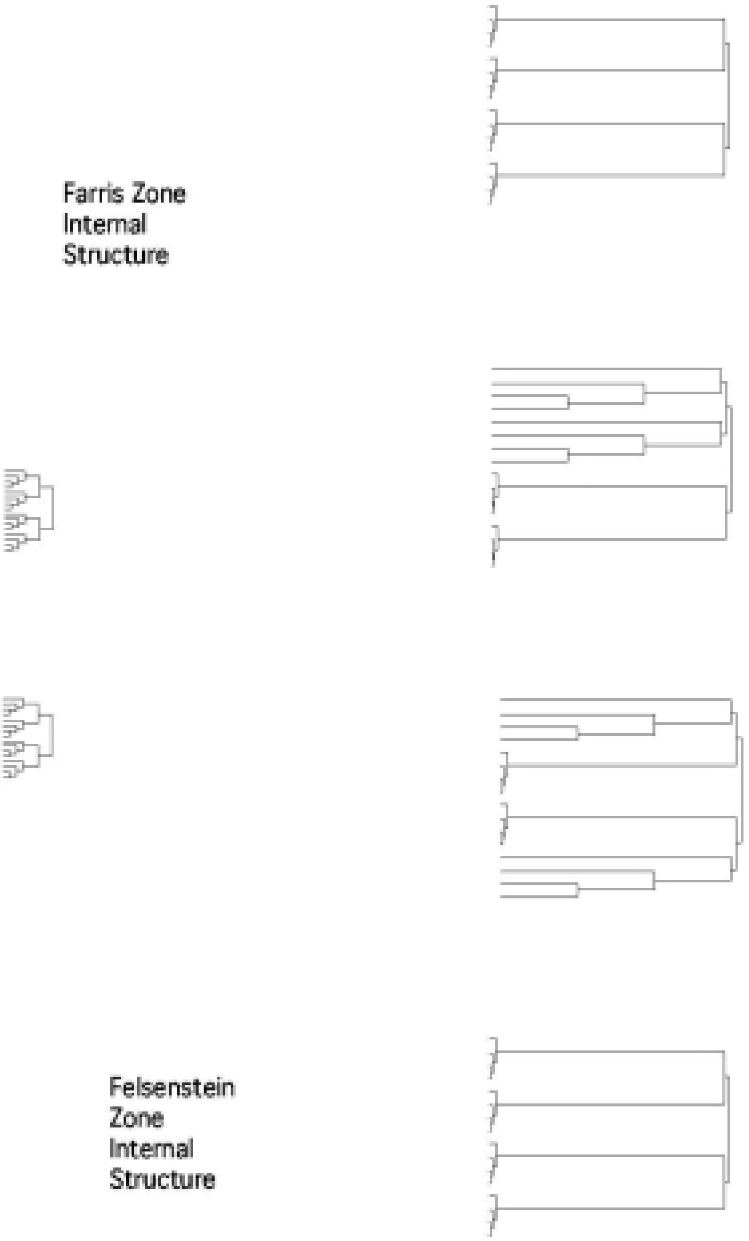

Farris Zone
Internal
Structure
Felsenstein
Zone
Internal
Structure

Figure 4.1

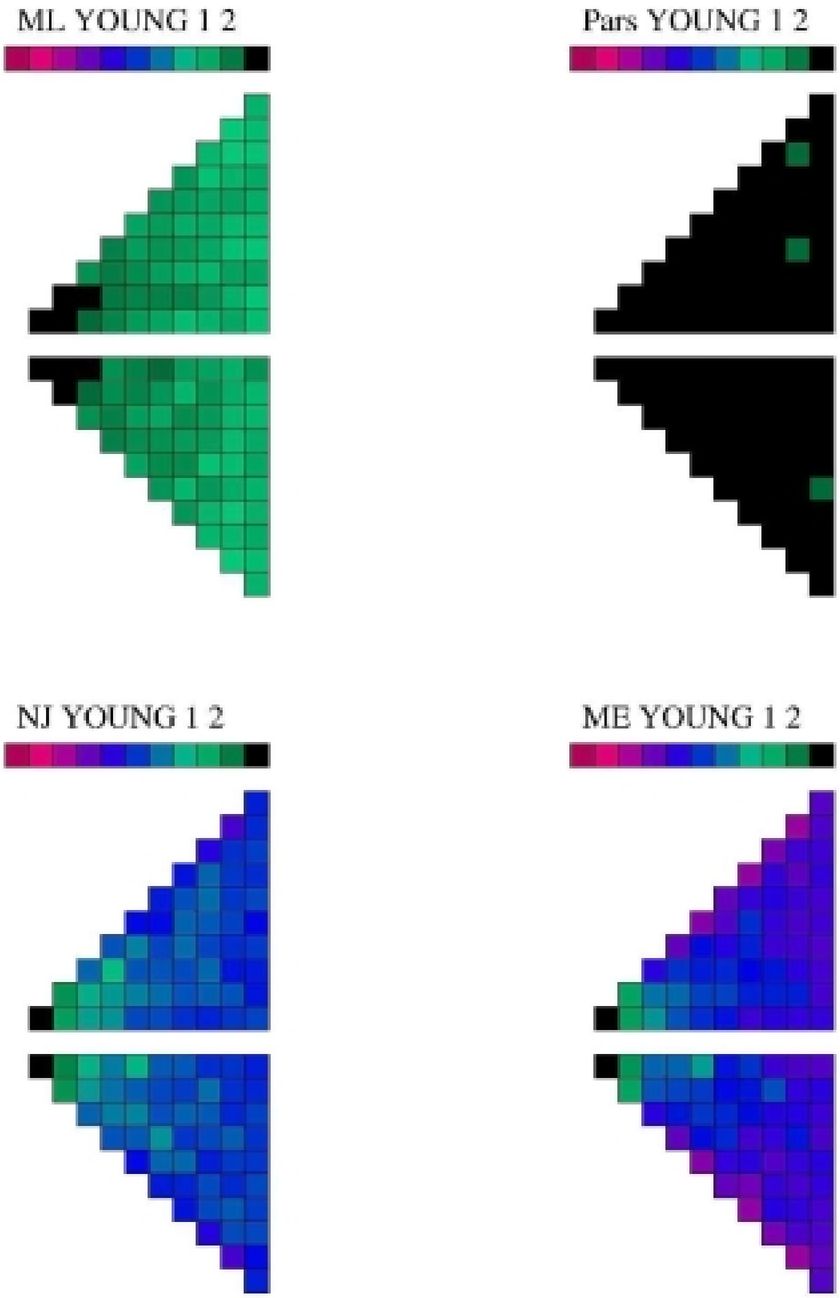


Figure 4.2

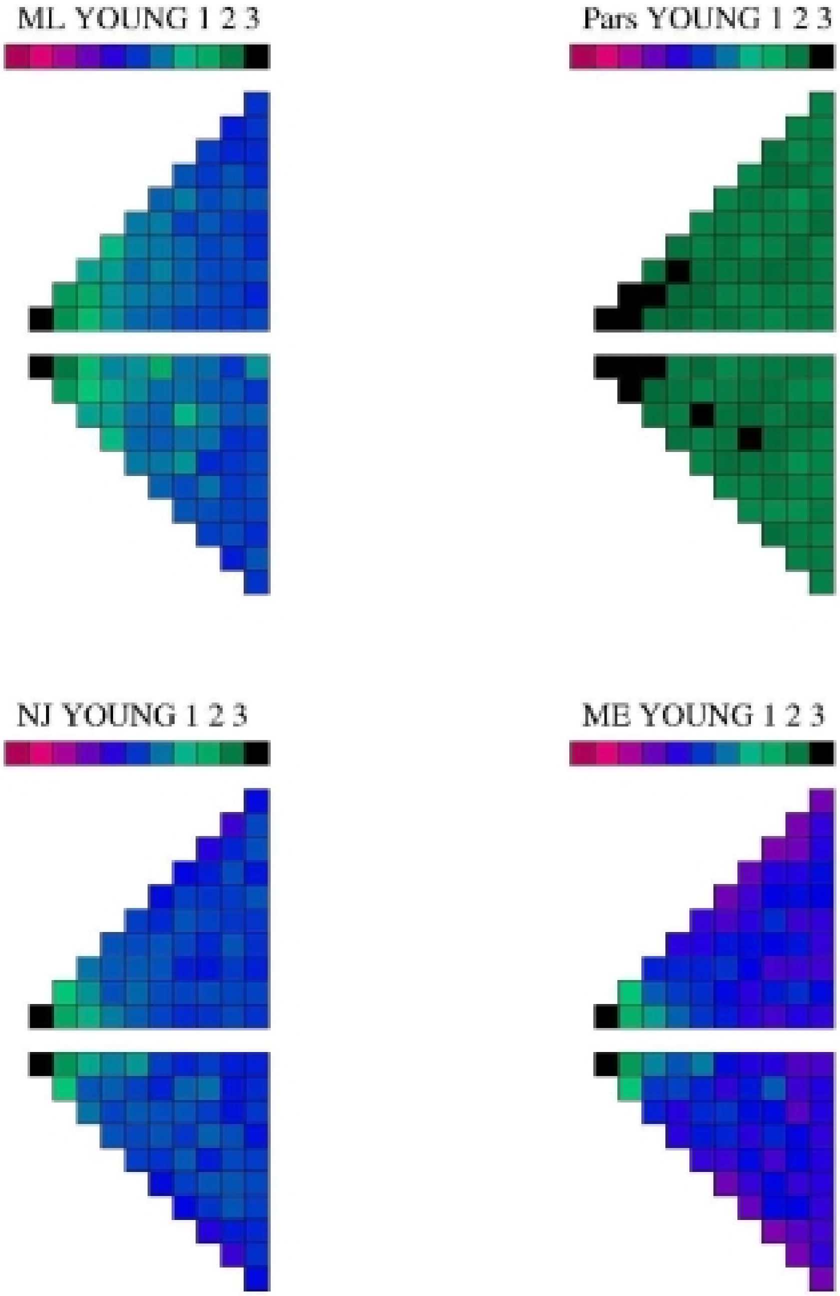


Figure 4.3

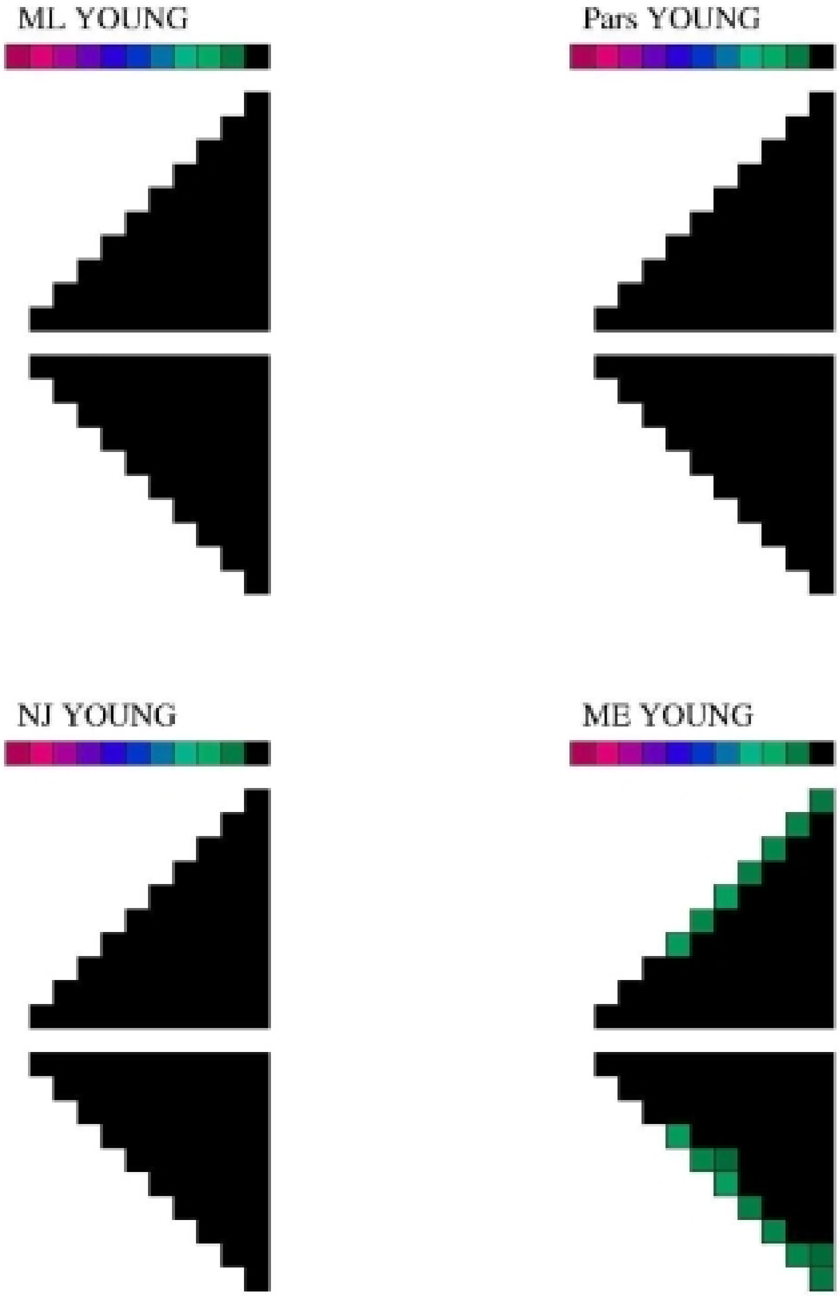


Figure 4.4

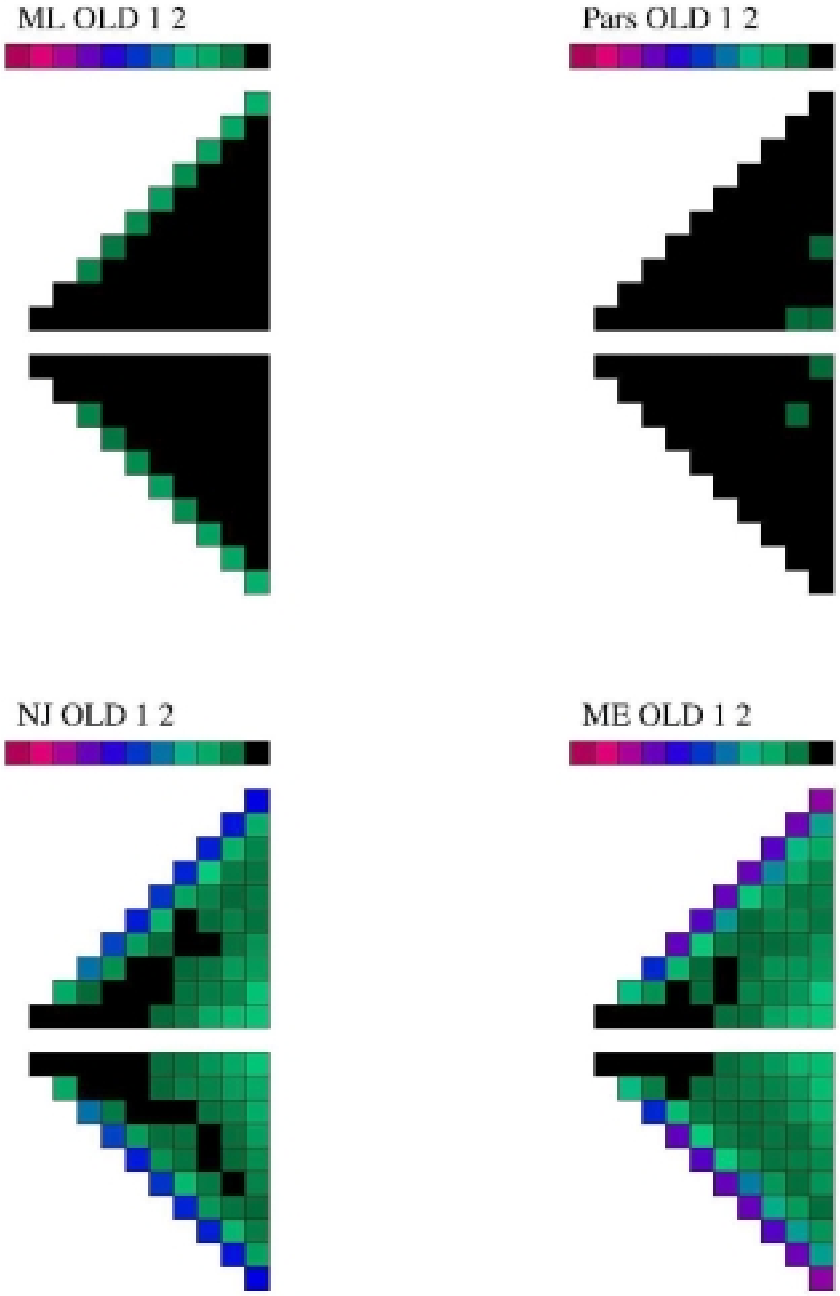


Figure 4.5

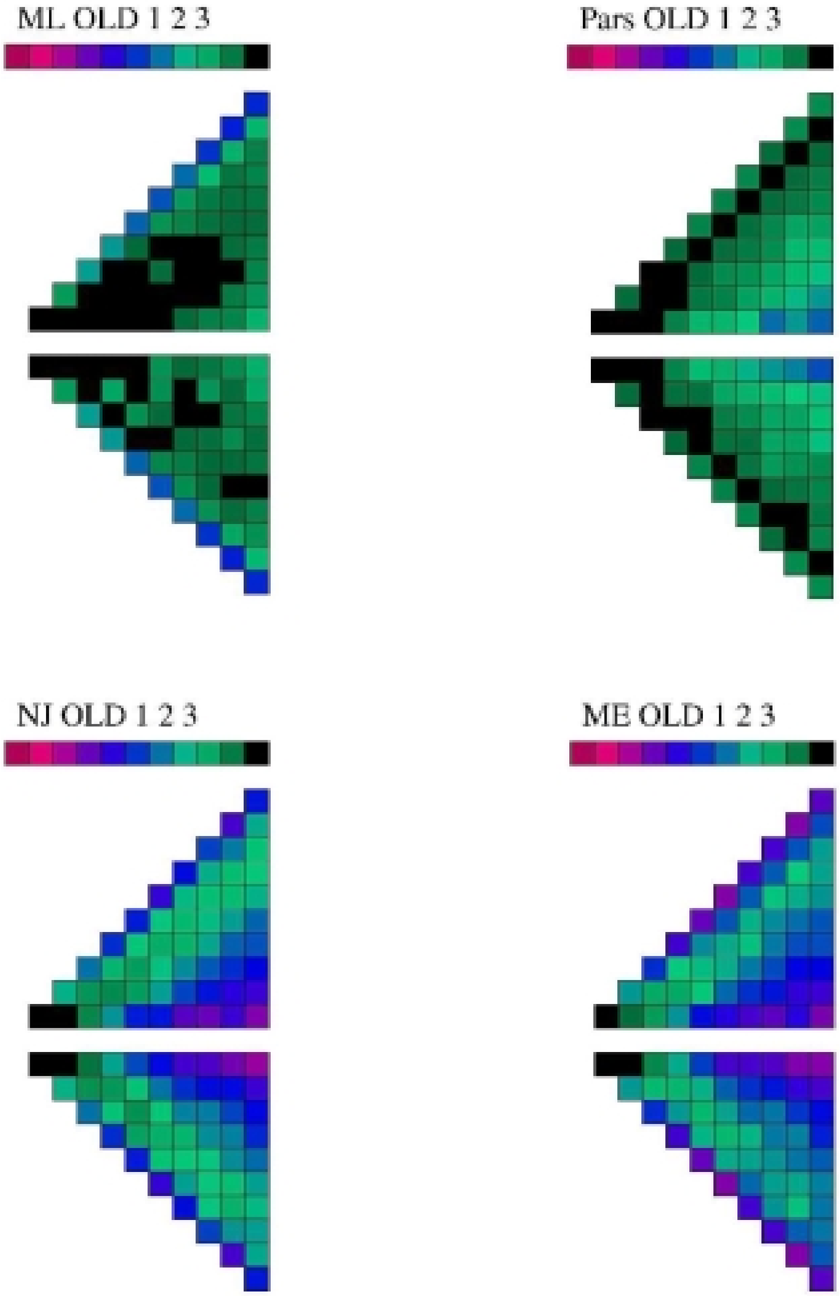


Figure 4.6

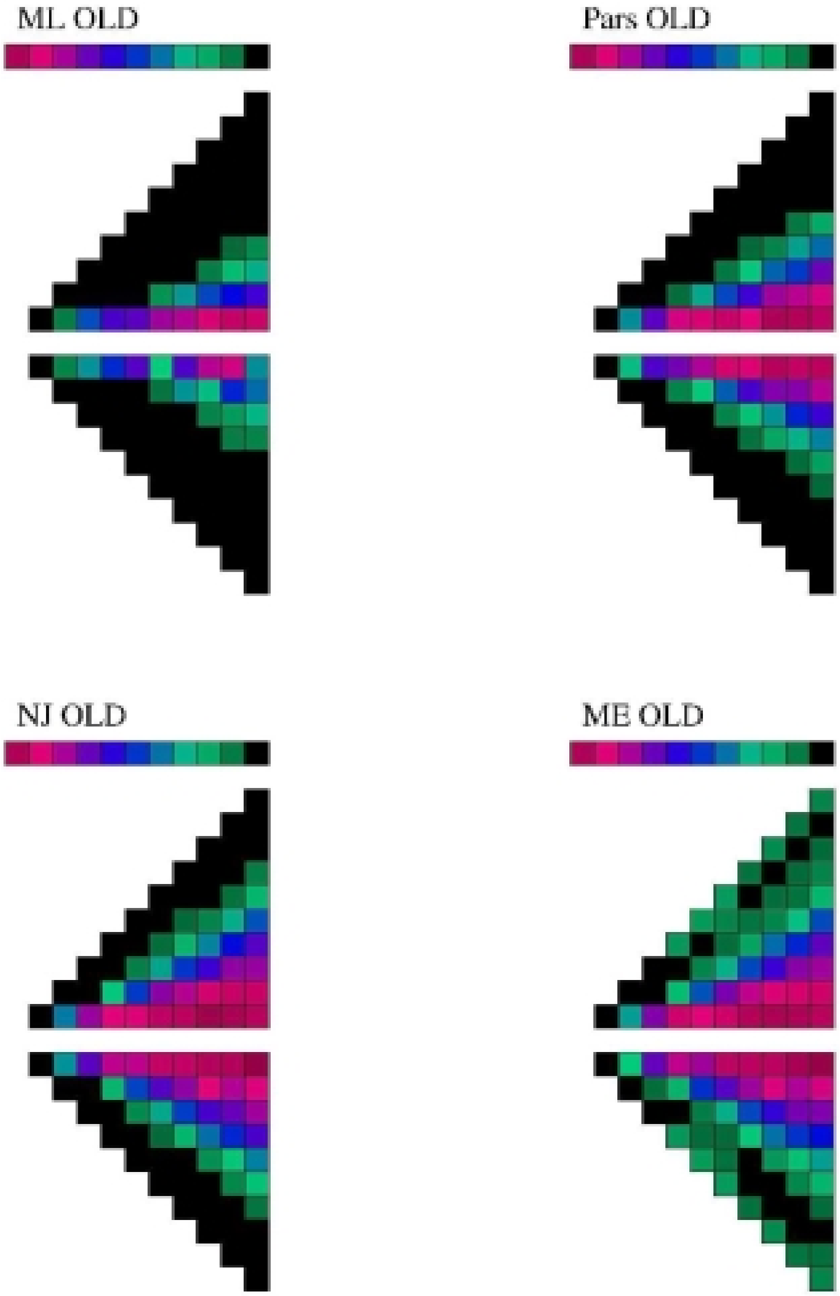


Figure 4.7

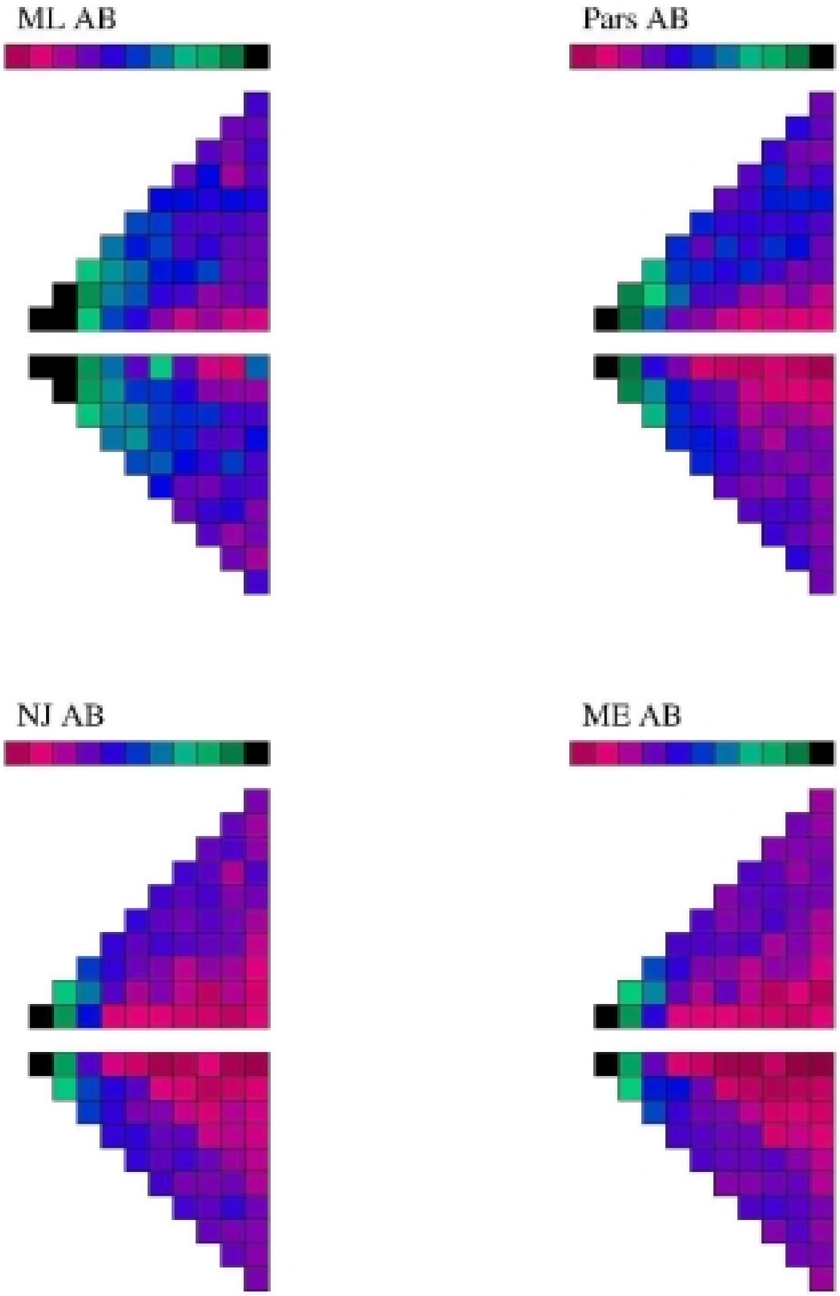


Figure 4.8

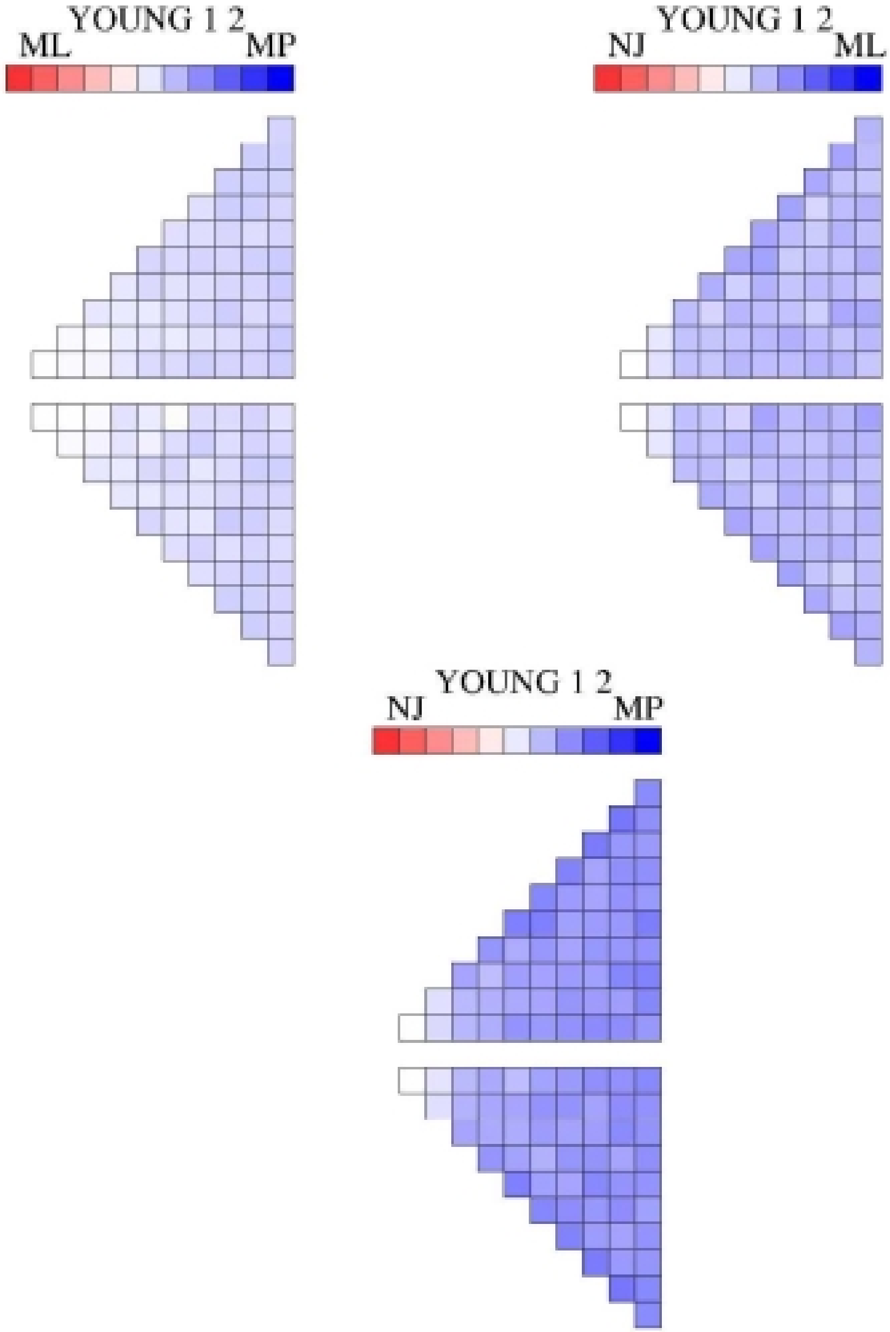


Figure 4.9

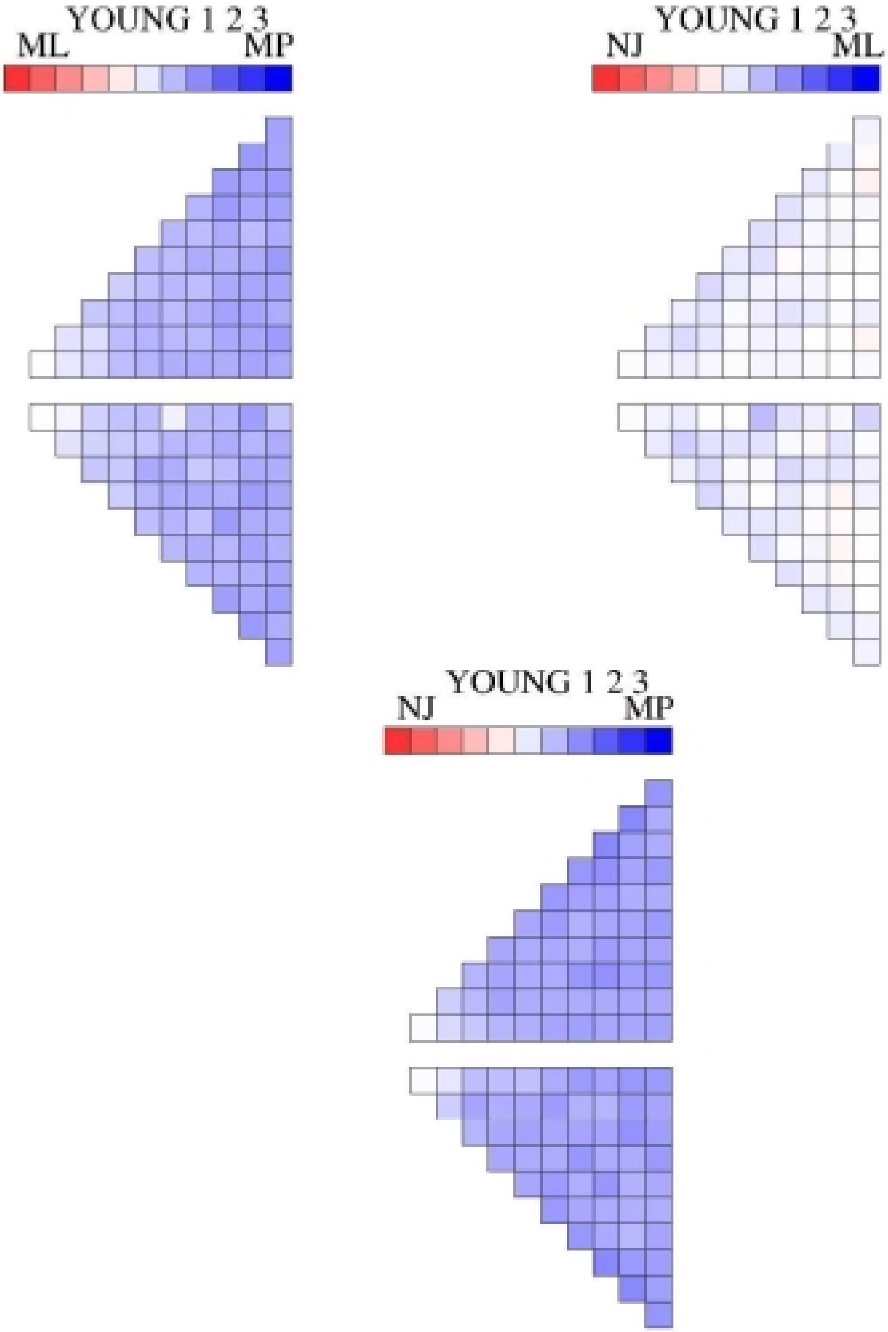


Figure 4.10

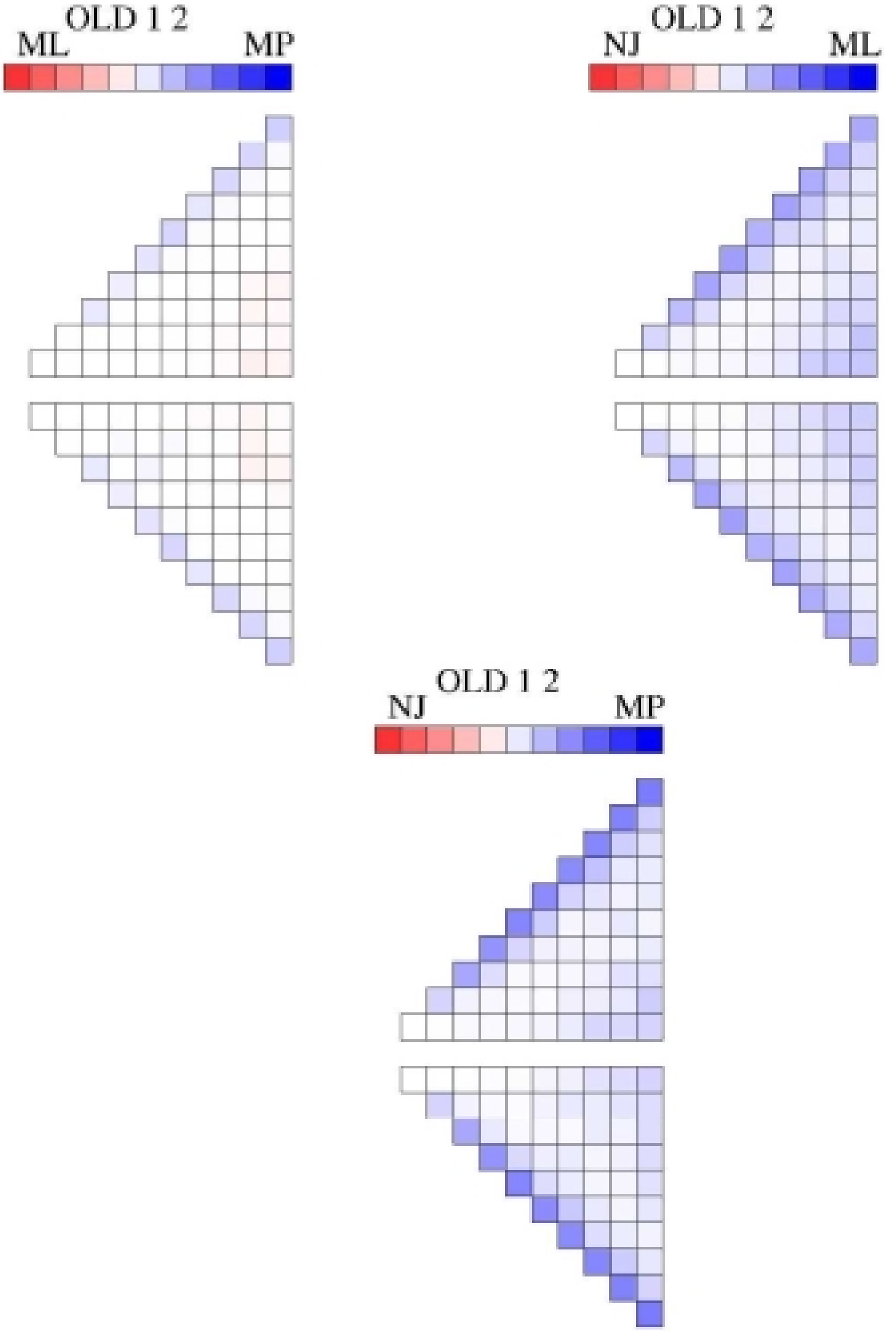


Figure 4.11

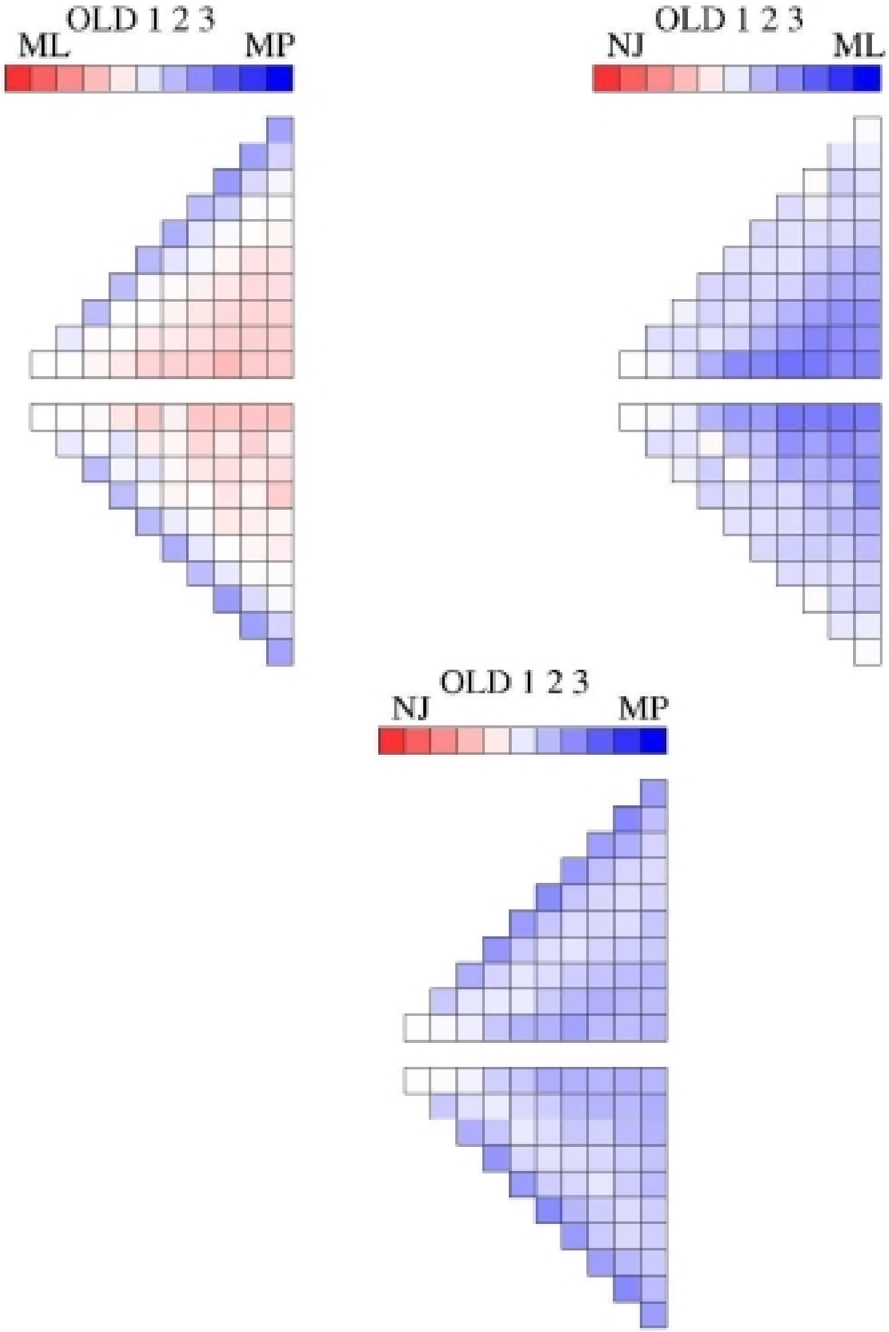


Figure 4.12

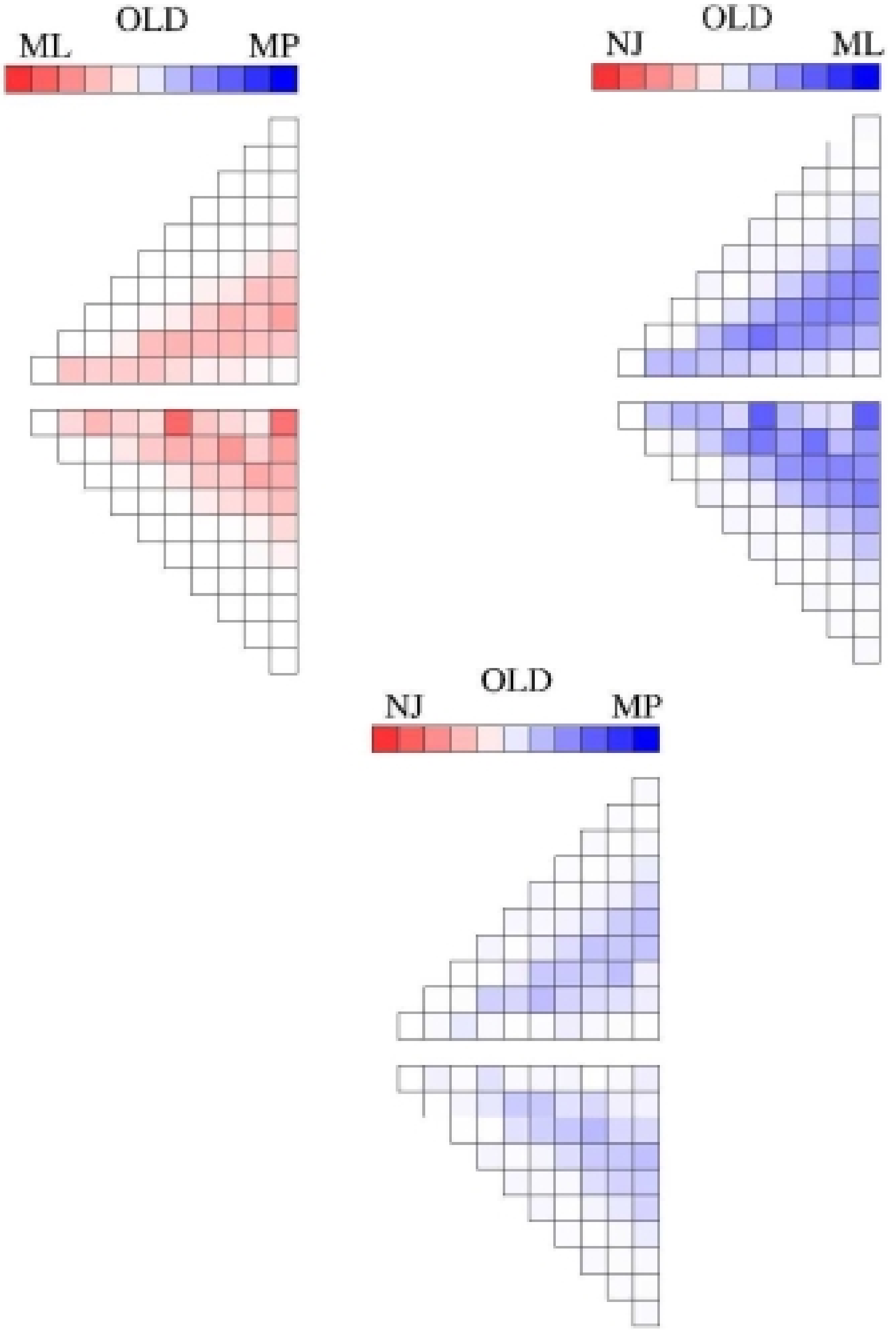


Figure 4.13

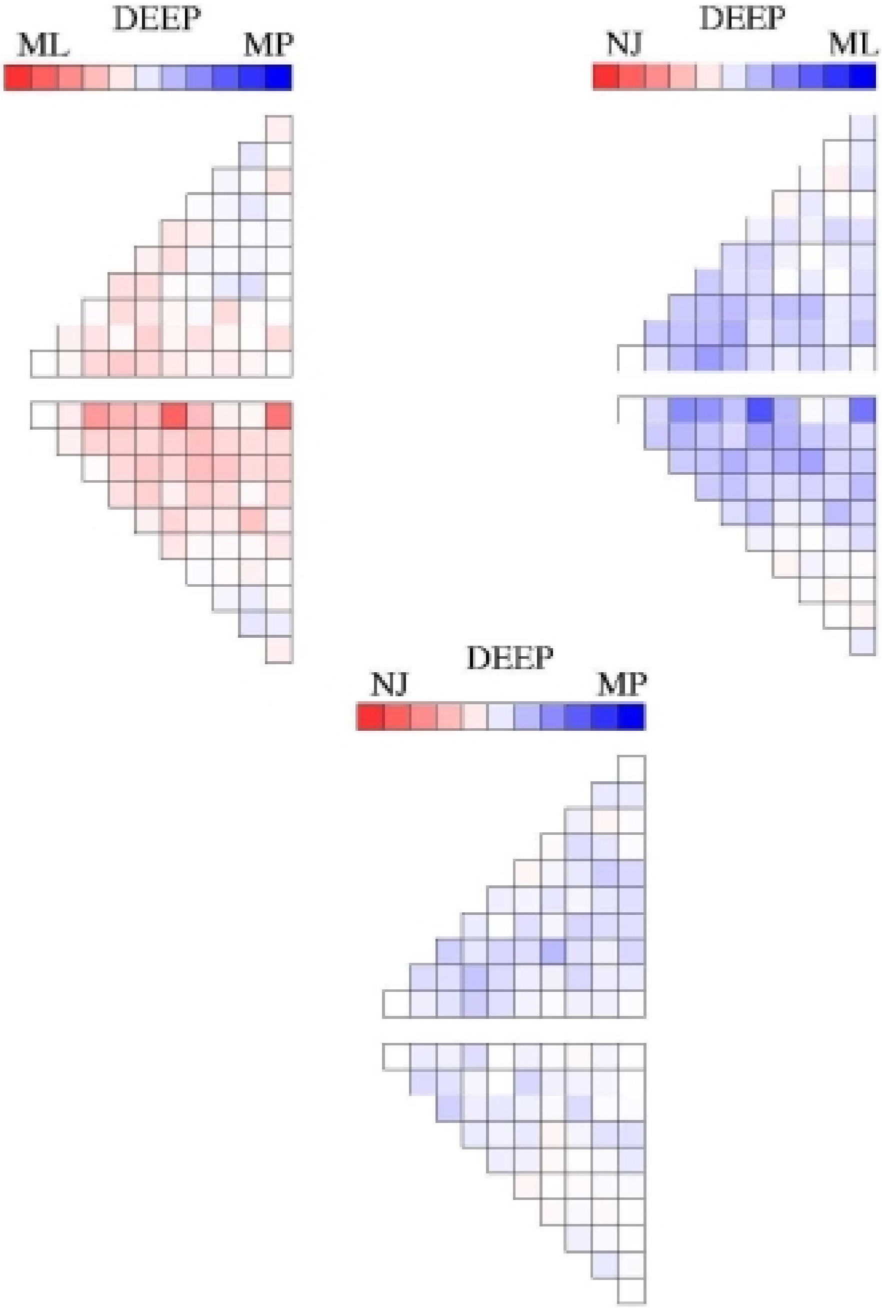


Figure 4.14

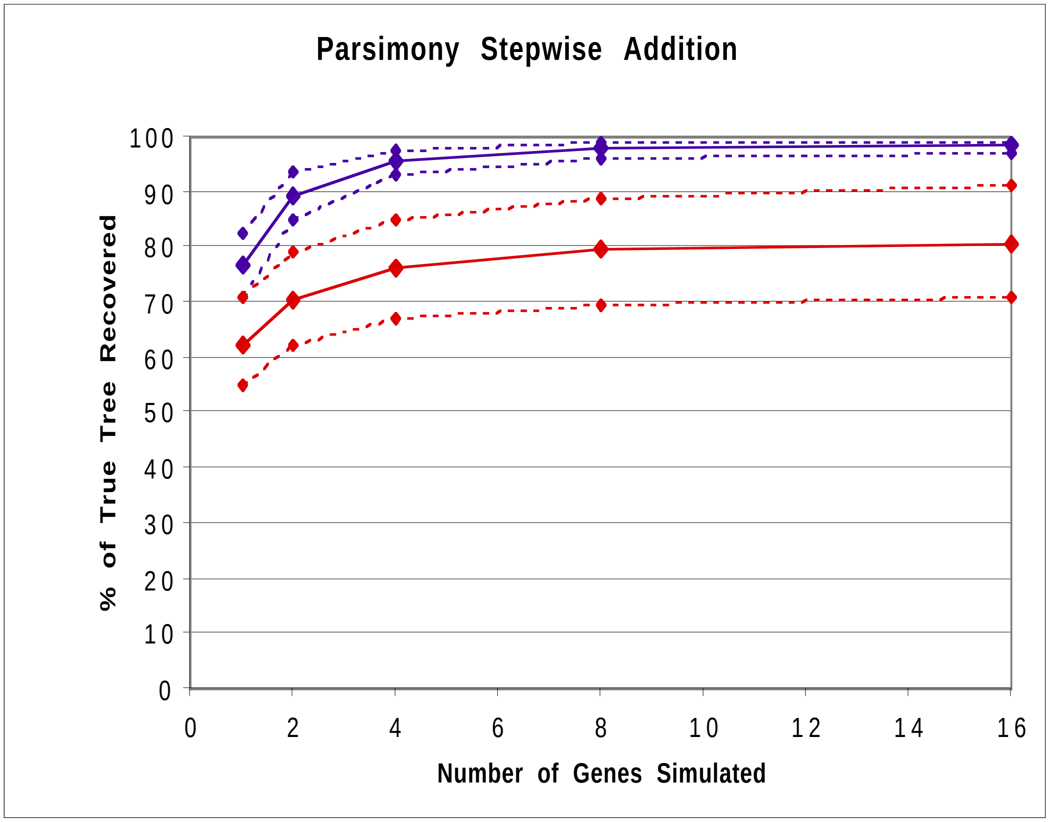


Figure 5.1

Performance of unordered-parsimony stepwise-addition on the short and long versions of the 228 taxon tree shown in red and blue respectively as a function of simulated sequence length (number of genes simulated). 95% confidence limits are shown as a dashed line. Performance is measured as the percentage of thetrue tree's internal branches present in the inferred tree.

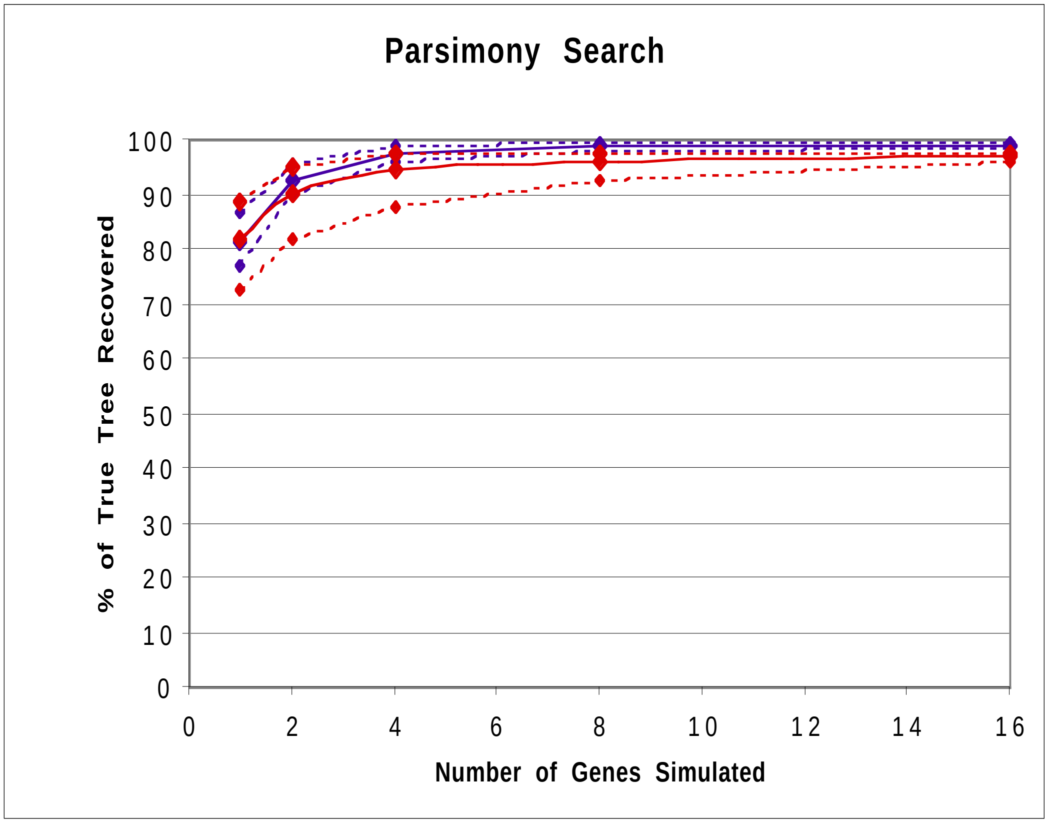


Figure 5.2

Performance of unordered-parsimony SPR-searches from a stepwise-addition tree on the short and long versions of the 228 taxon tree shown in red and blue respectively as a function of simulated sequence length (number of genes simulated). 95% confidence limits are shown as a dashed line. Performance is measured as the percentage of thetrue tree's internal branches present in the inferred tree.

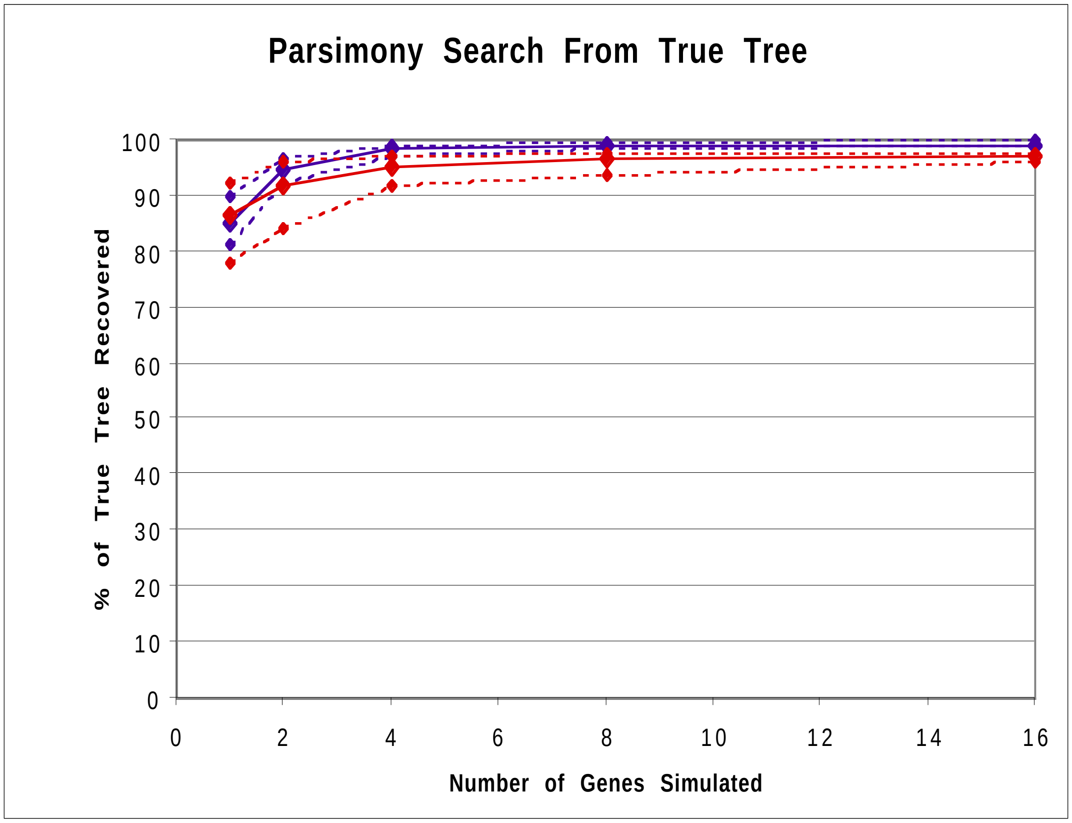


Figure 5.3

Performance of unordered-parsimony SPR-searches from the true tree on the short and long versions of the 228 taxon tree shown in red and blue respectively as a function of simulated sequence length (number of genes simulated). 95% confidence limits are shown as a dashed line. Performance is measured as the percentage of thetrue tree's internal branches present in the inferred tree.

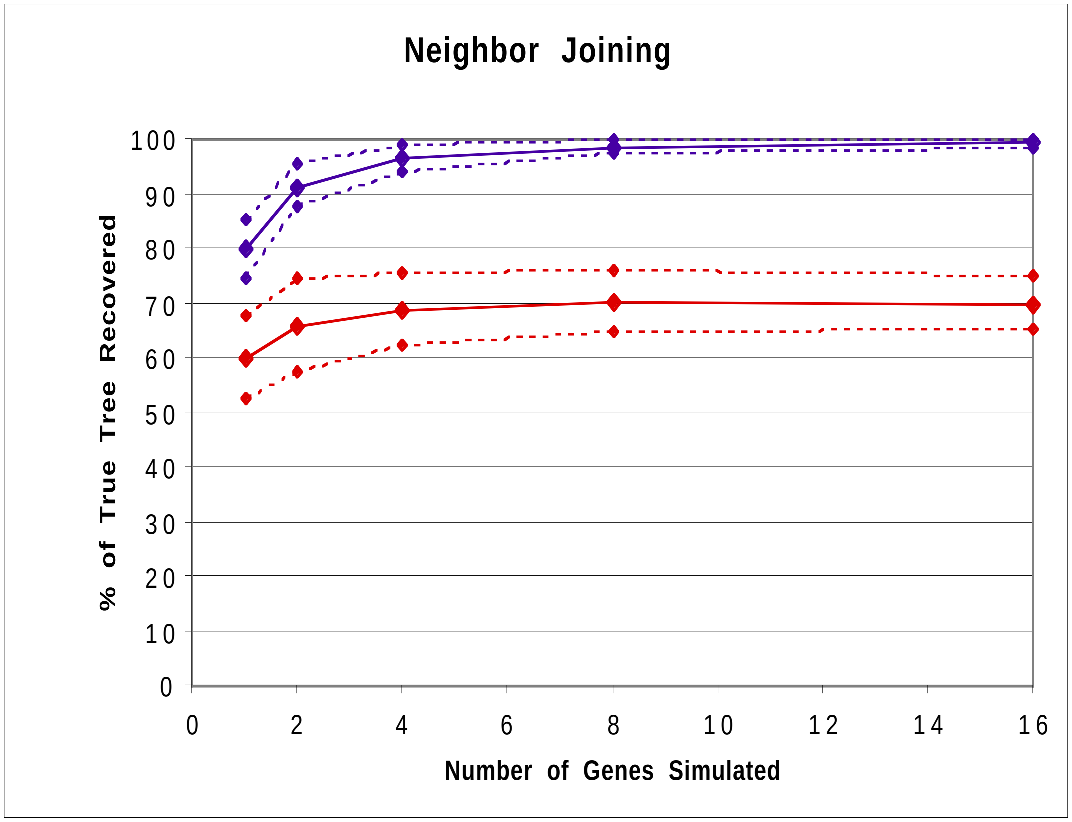


Figure 5.4

Performance of neighbor joining on the short and long versions of the 228 taxon tree shown in red and blue respectively as a function of simulated sequence length (number of genes simulated). 95% confidence limits are shown as a dashed line. Performance is measured as the percentage of thetrue tree's internal branches present in the inferred tree.

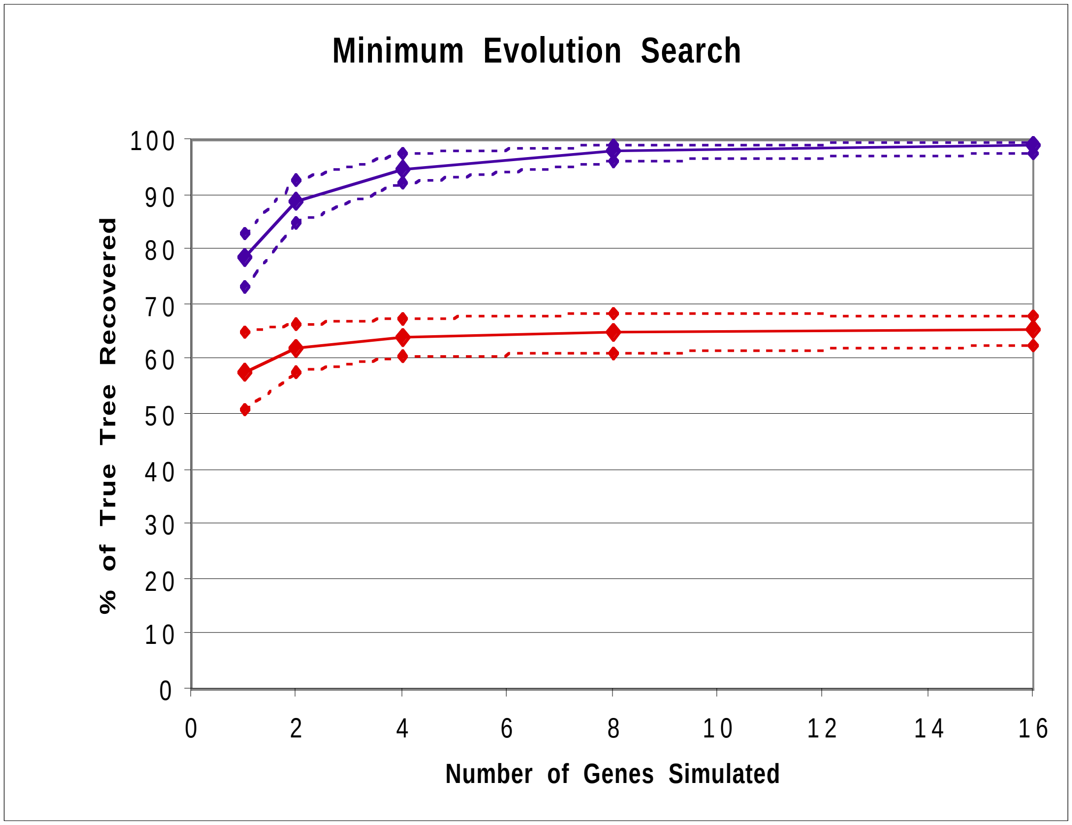


Figure 5.5

Performance of minimum evolution SPR-searches from a neighbor joining tree on the short and long versions of the 228 taxon tree shown in red and blue respectively as a function of simulated sequence length (number of genes simulated). 95% confidence limits are shown as a dashed line. Performance is measured as the percentage of thetrue tree's internal branches present in the inferred tree.

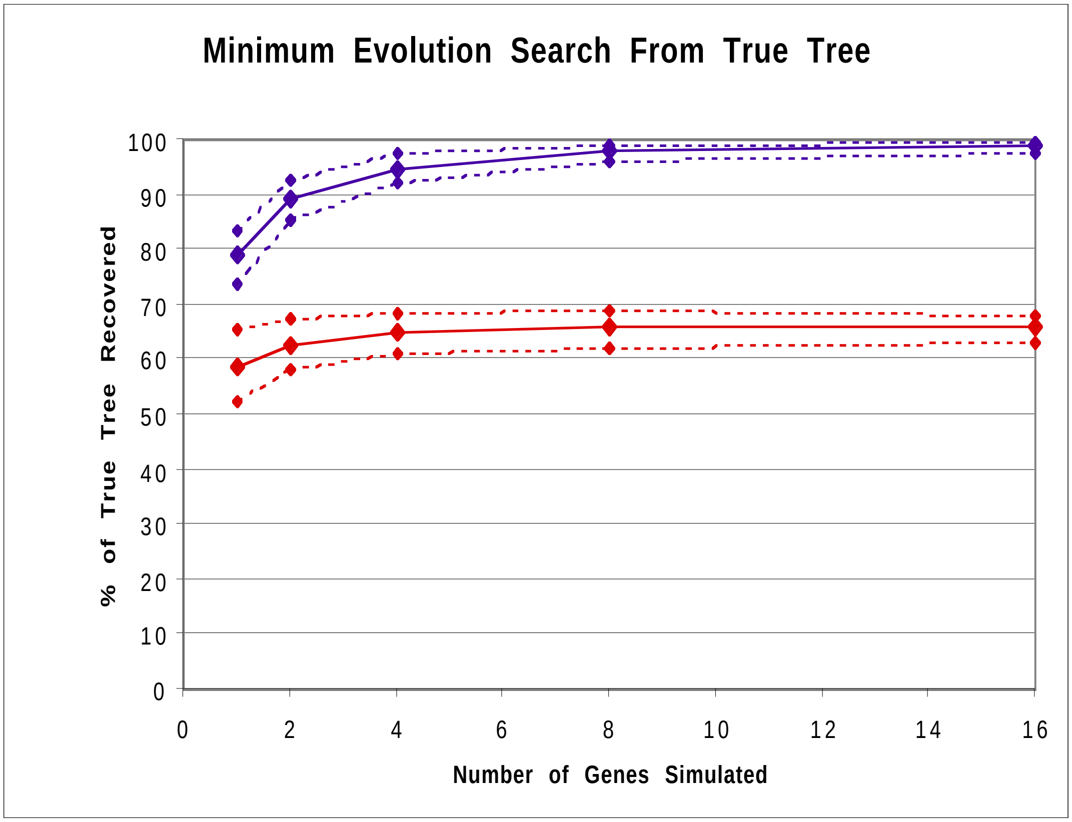


Figure 5.6

Performance of minimum evolution SPR-searches from the true tree on the short and long versions of the 228 taxon tree shown in red and blue respectively as a function of simulated sequence length (number of genes simulated). 95% confidence limits are shown as a dashed line.

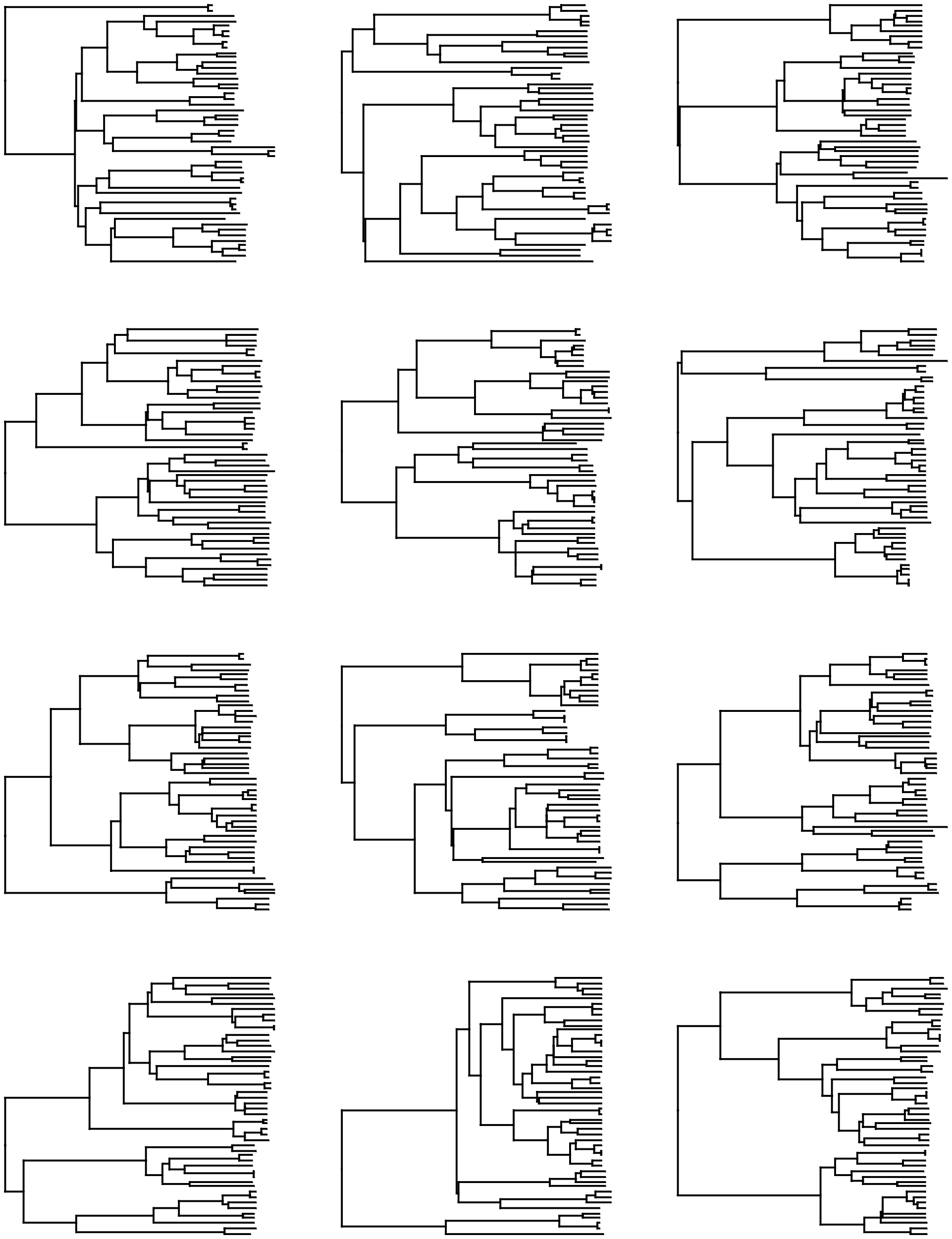


Figure 6.1

Twelve examples of 50 taxon tree shapes (scale is not constant) generated by the modified Yule process using scaling parameters = 0.5 and rho=0.1.

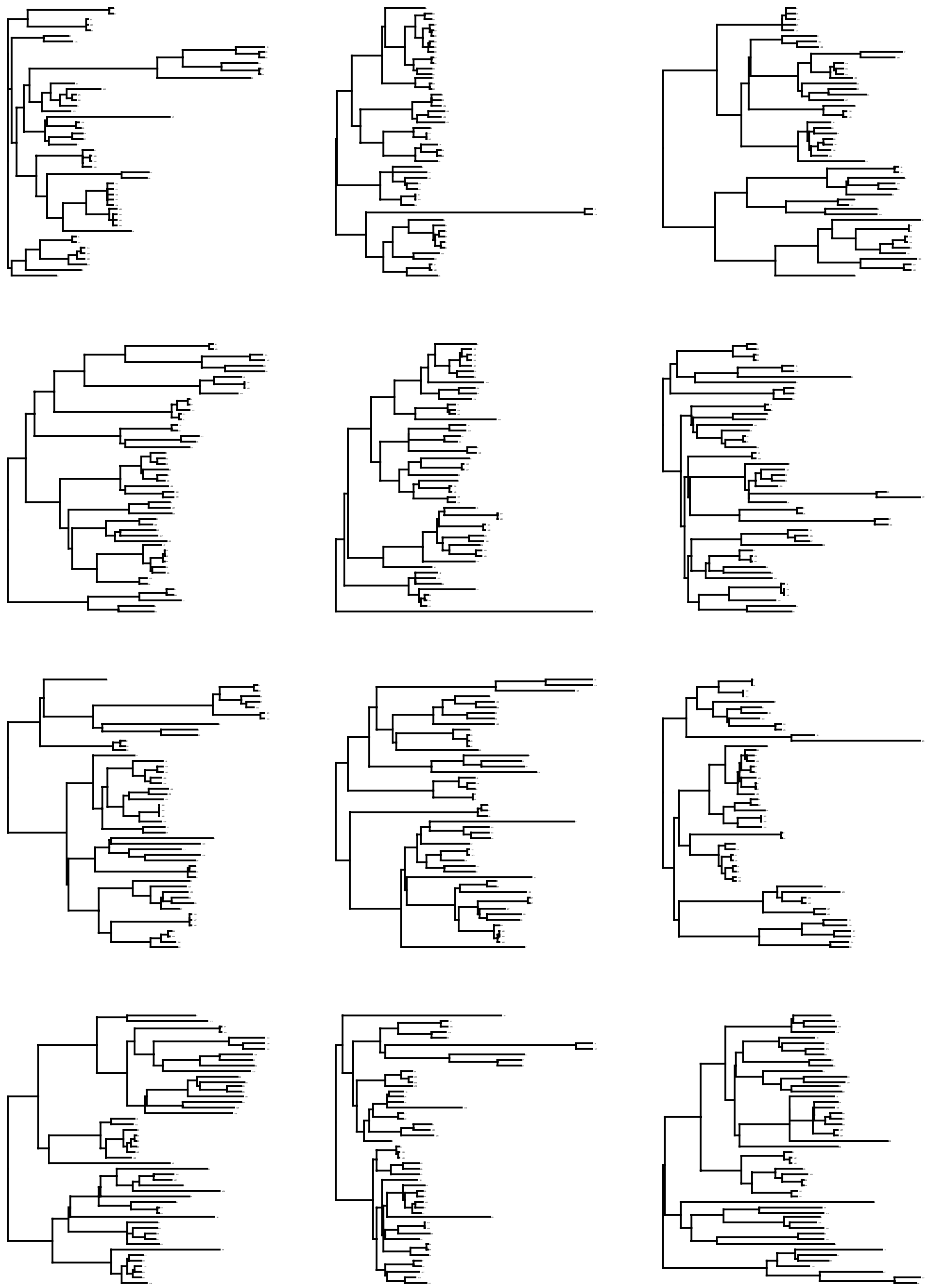


Figure 6.2

Twelve examples of 50 taxon tree shapes (scale is not constant) generated by the modified Yule process using scaling parameters = 0.5 and rho=1.0.

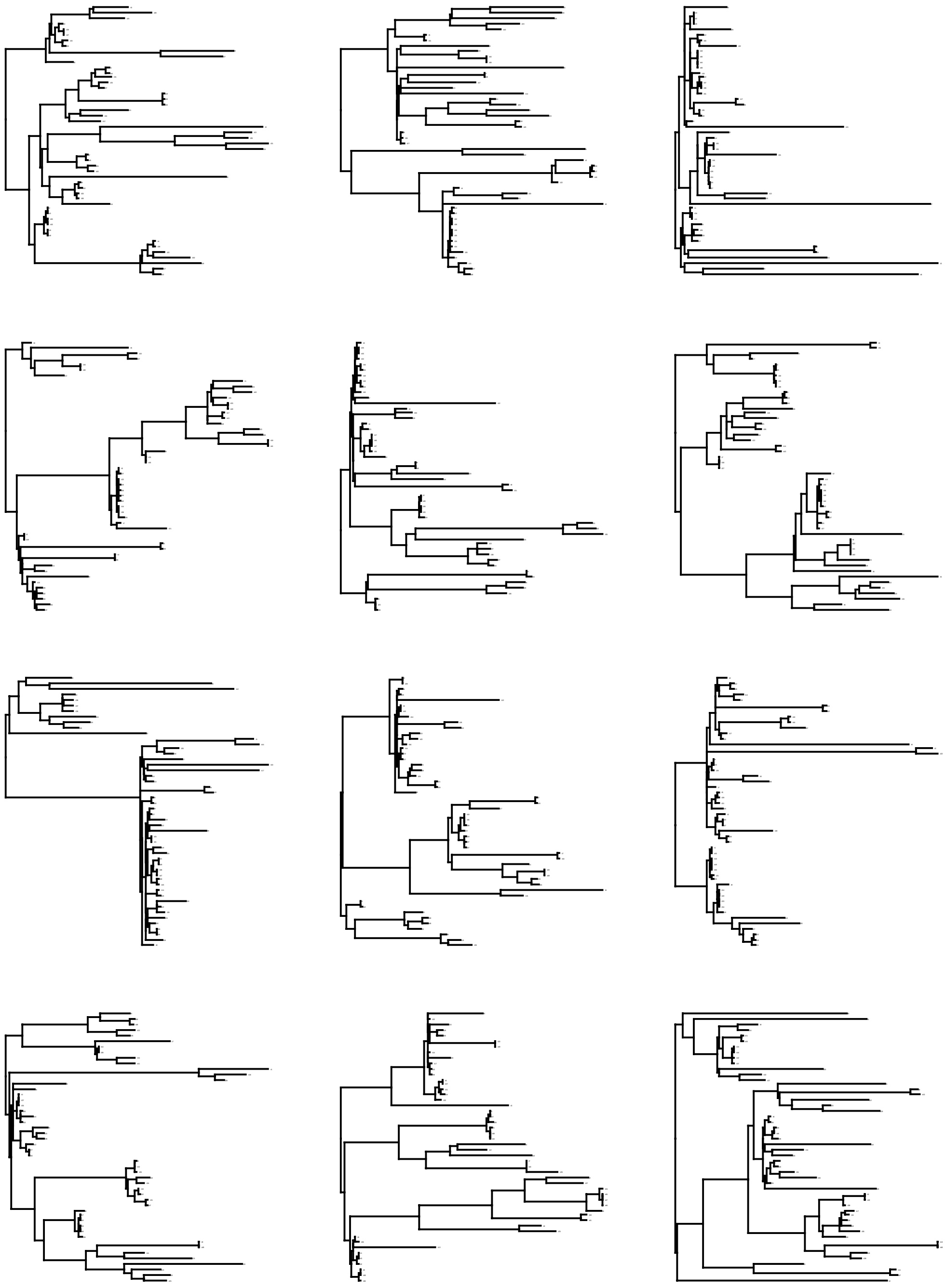


Figure 6.3

Twelve examples of 50 taxon tree shapes (scale is not constant) generated by the modified Yule process using scaling parameters = 0.5 and rho=10.0.

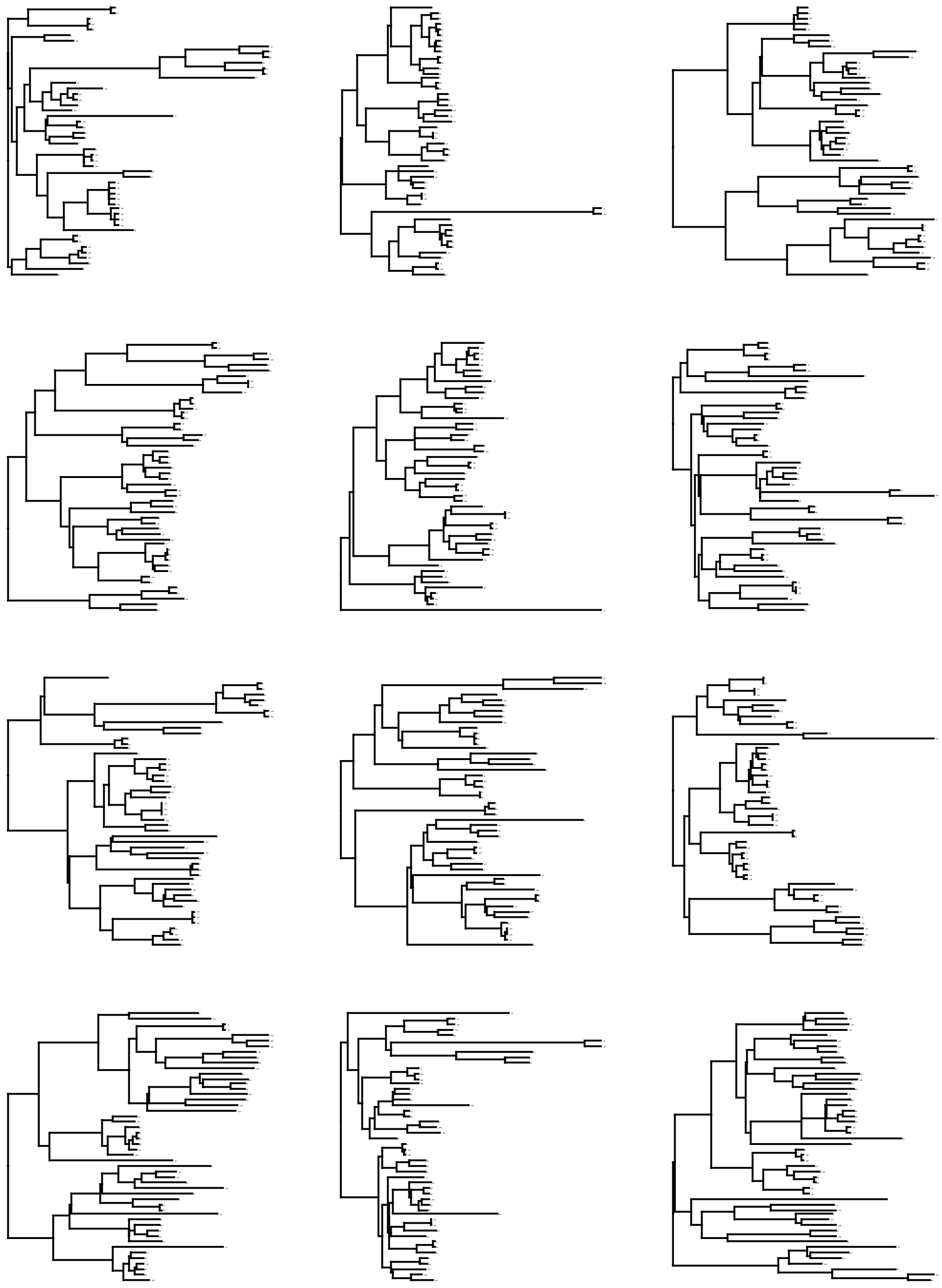


Figure 6.4

Twelve examples of 50 taxon tree shapes (scale is not constant) generated by the modified Yule process using scaling parameters = 1.0 and rho=0.1.

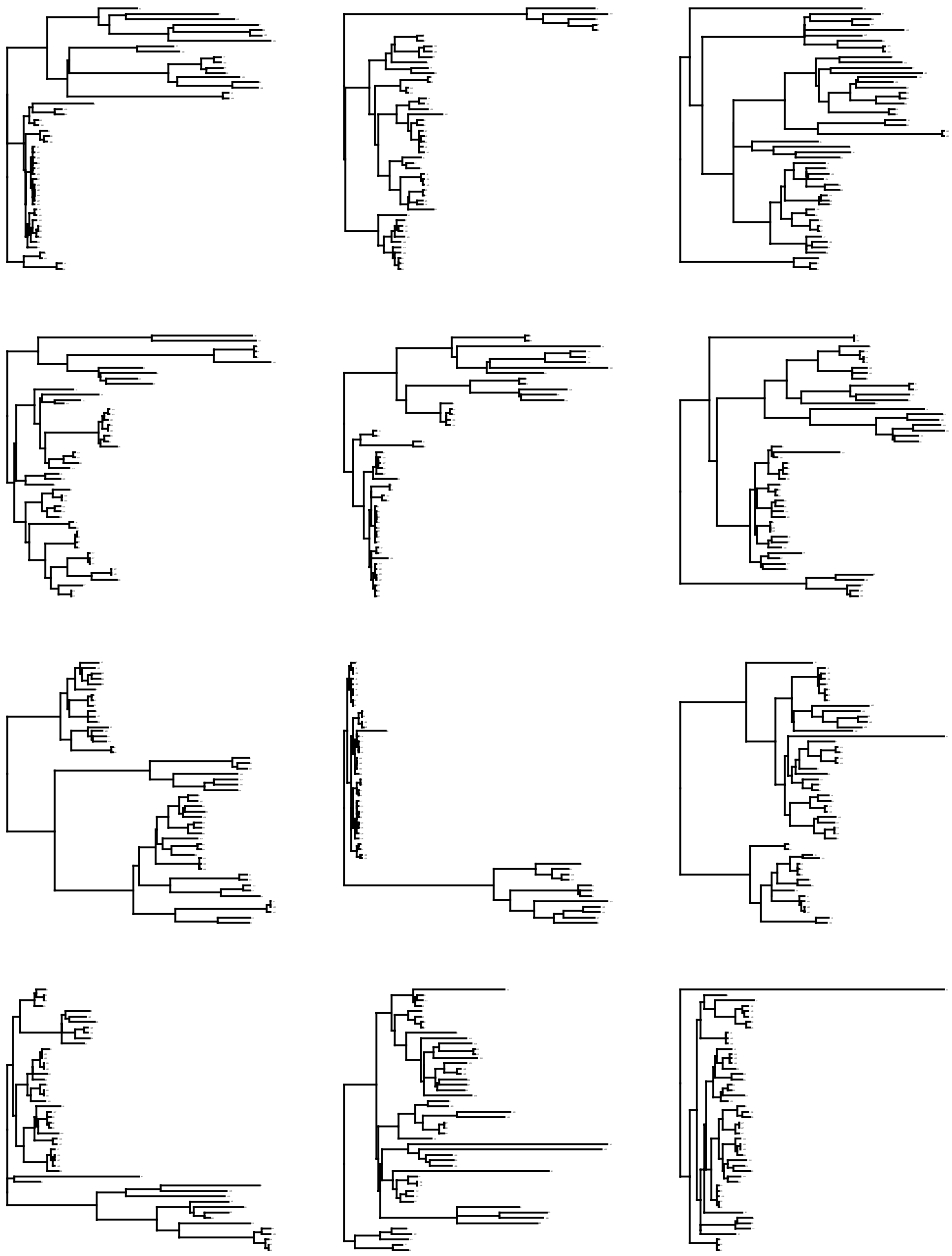


Figure 6.5

Twelve examples of 50 taxon tree shapes (scale is not constant) generated by the modified Yule process using scaling parameters = 1.0 and rho=1.0.

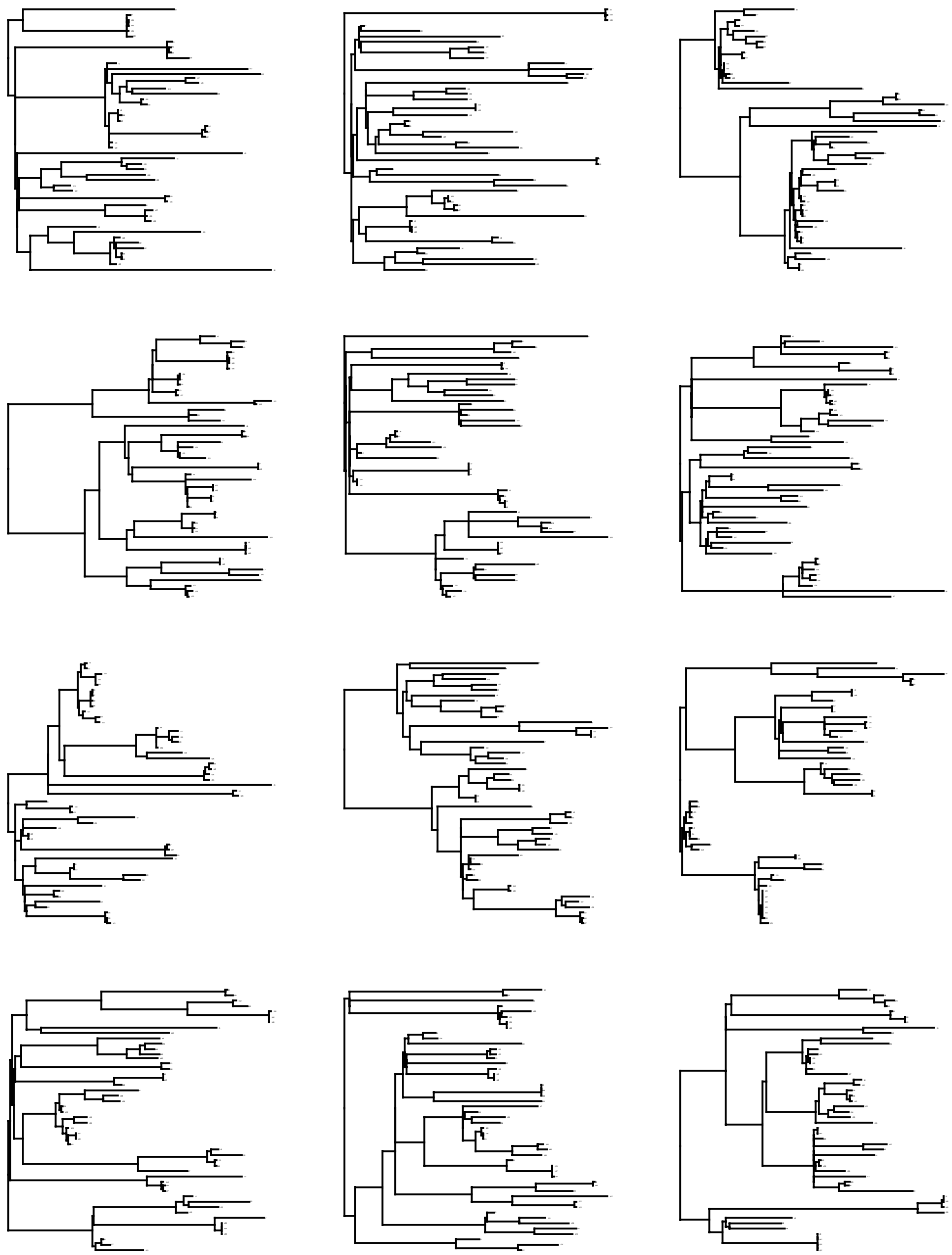


Figure 6.6

Twelve examples of 50 taxon tree shapes (scale is not constant) generated by the modified Yule process using scaling parameters = 1.0 and rho=10.0

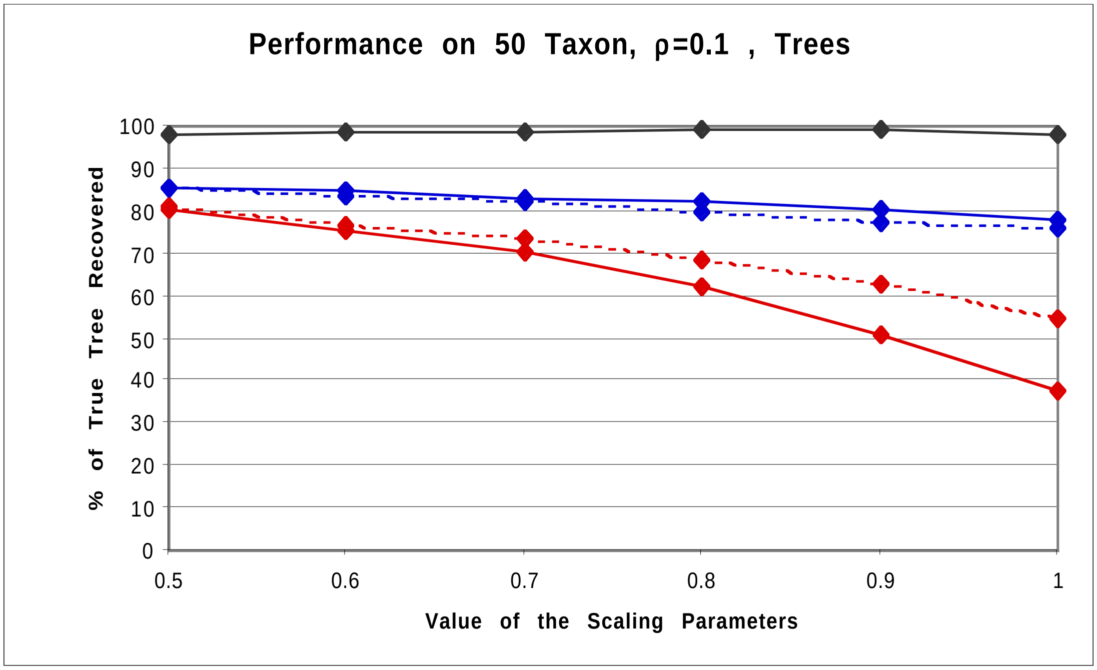


Figure 6.7

Performance of parsimony searches (solid blue), parsimony stepwise addition (dashed blue), neighbor joining (dashed red) and minimum evolution (solid red) on 50-taxon trees generated under with rho=0.1 and plotted over all 6 scaling parameter settings. Performance is measured as the percentage of thetrue tree’s internal branches present in the inferred tree. The black line indicates the best possible performance.

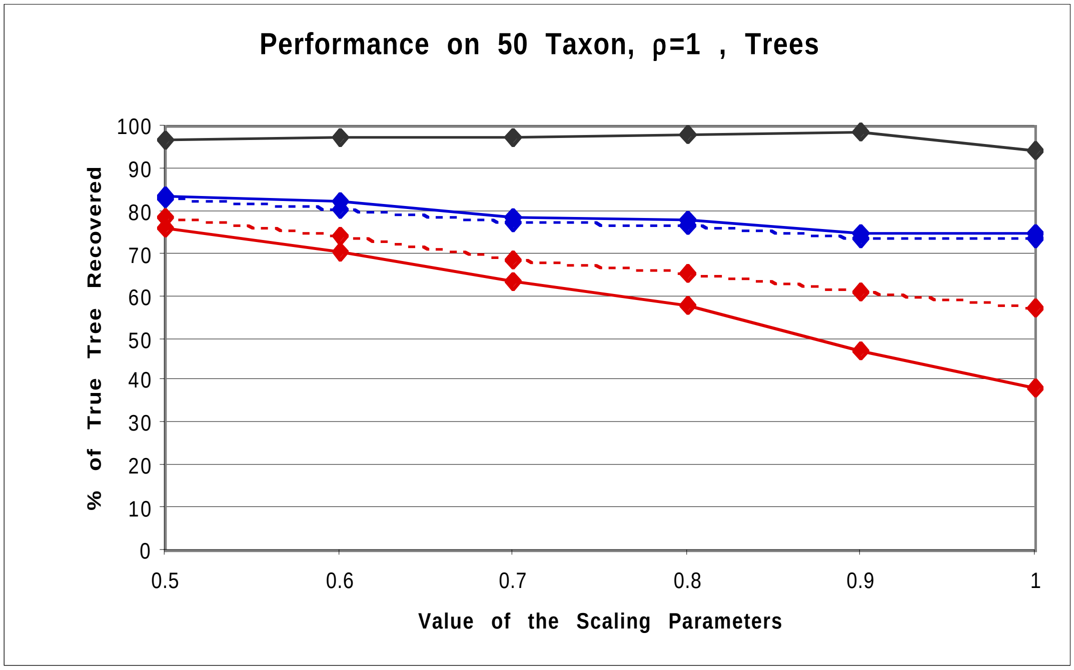


Figure 6.8

Performance of parsimony searches (solid blue), parsimony stepwise addition (dashed blue), neighbor joining (dashed red) and minimum evolution (solid red) on 50-taxon trees generated under with rho=1 and plotted over all 6 scaling parameter settings. Performance is measured as the percentage of thetrue tree’s internal branches present in the inferred tree. The black line indicates the best possible performance.

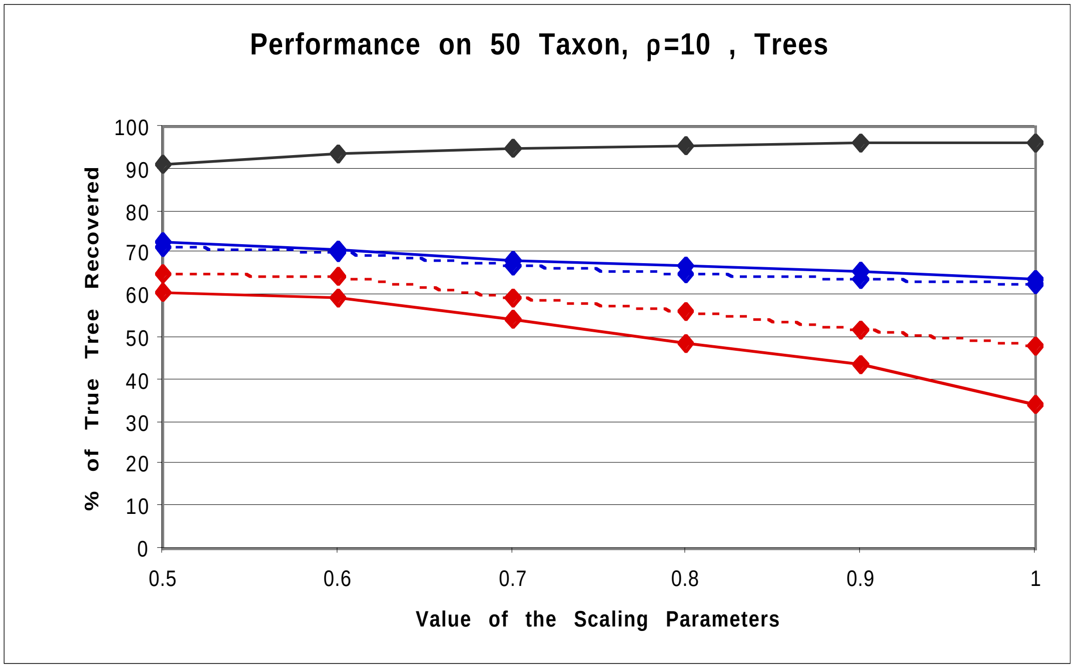


Figure 6.9

Performance of parsimony searches (solid blue), parsimony stepwise addition (dashed blue), neighbor joining (dashed red) and minimum evolution (solid red) on 50-taxon trees generated under with rho=10 and plotted over all 6 scaling parameter settings. Performance is measured as the percentage of thetrue tree's internal branches present in the inferred tree. The black line indicates the best possible performance.

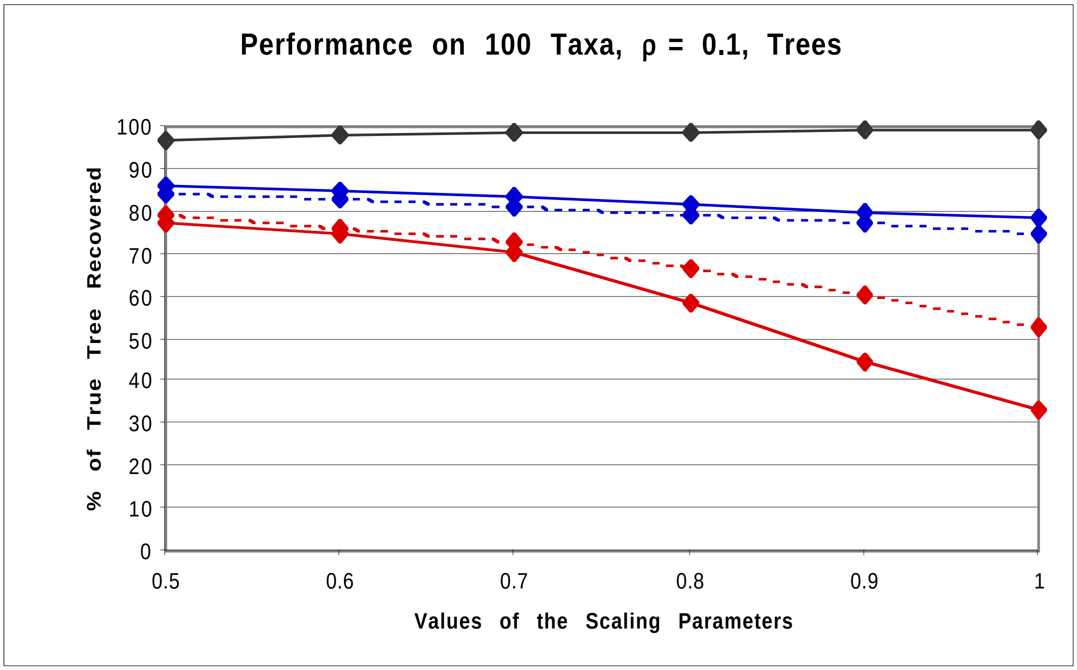


Figure 6.10

Performance of parsimony searches (solid blue), parsimony stepwise addition (dashed blue), neighbor joining (dashed red) and minimum evolution (solid red) on 100-taxon trees generated under with rho=0.1 and plotted over all 6 scaling parameter settings. Performance is measured as the percentage of thetrue tree's internal branches present in the inferred tree. The black line indicates the best possible performance.

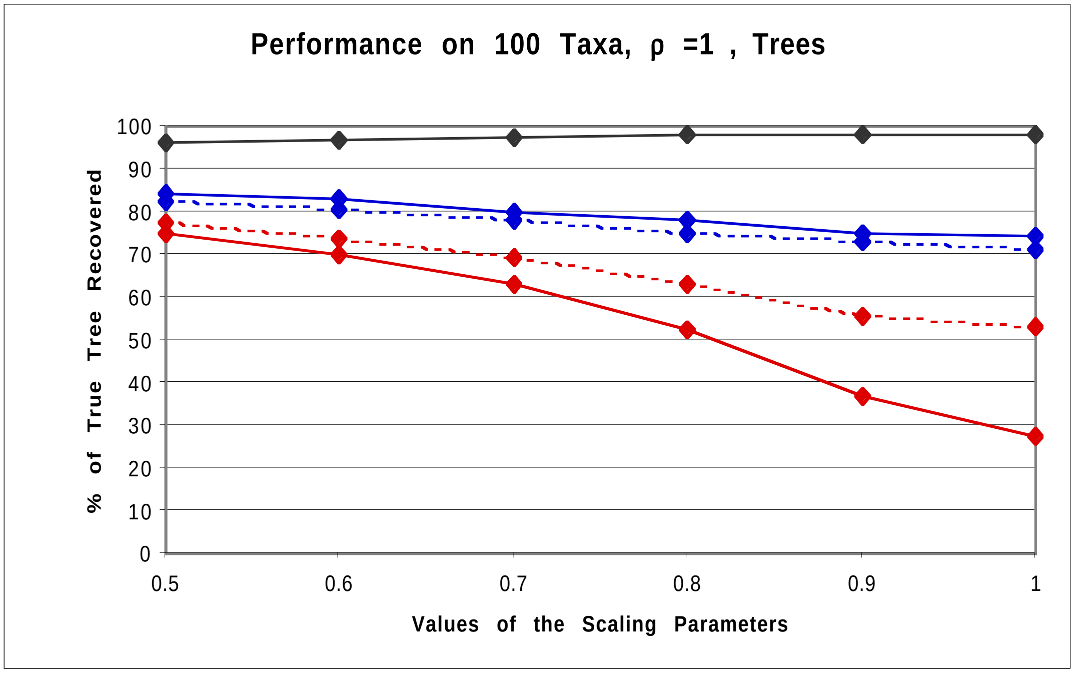


Figure 6.11

Performance of parsimony searches (solid blue), parsimony stepwise addition (dashed blue), neighbor joining (dashed red) and minimum evolution (solid red) on 100-taxon trees generated under with rho=1 and plotted over all 6 scaling parameter settings. Performance is measured as the percentage of thetrue tree’s internal branches present in the inferred tree. The black line indicates the best possible performance.

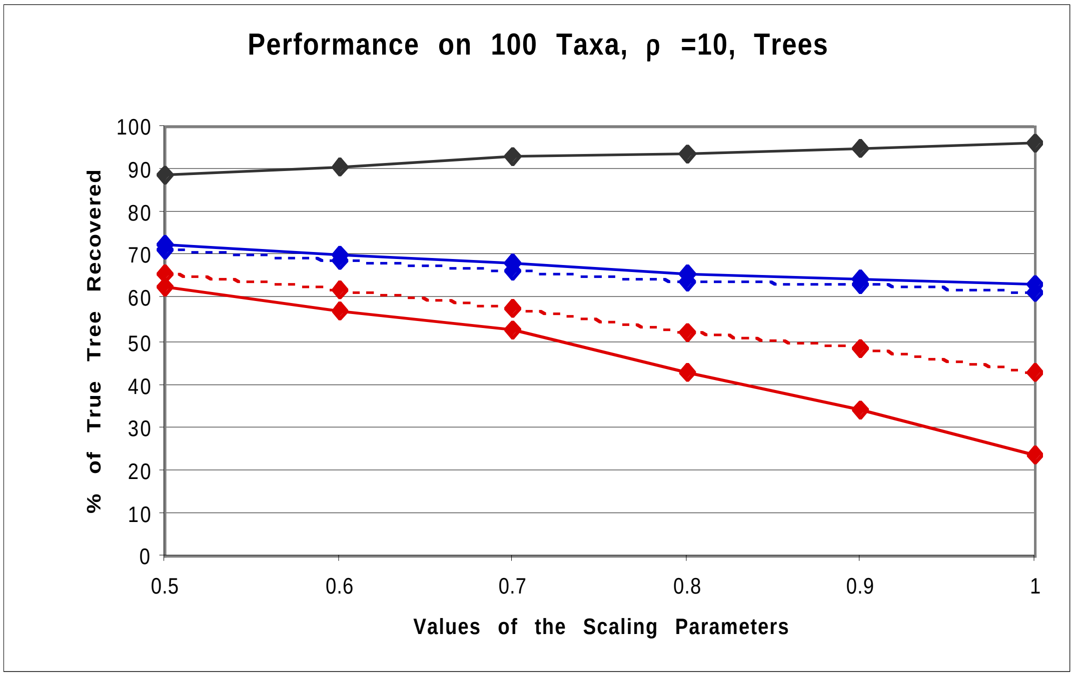


Figure 6.12

Performance of parsimony searches (solid blue), parsimony stepwise addition (dashed blue), neighbor joining (dashed red) and minimum evolution (solid red) on 100-taxon trees generated under with rho=0.1 and plotted over all 6 scaling parameter settings. Performance is measured as the percentage of thetrue tree's internal branches present in the inferred tree. The black line indicates the best possible performance.

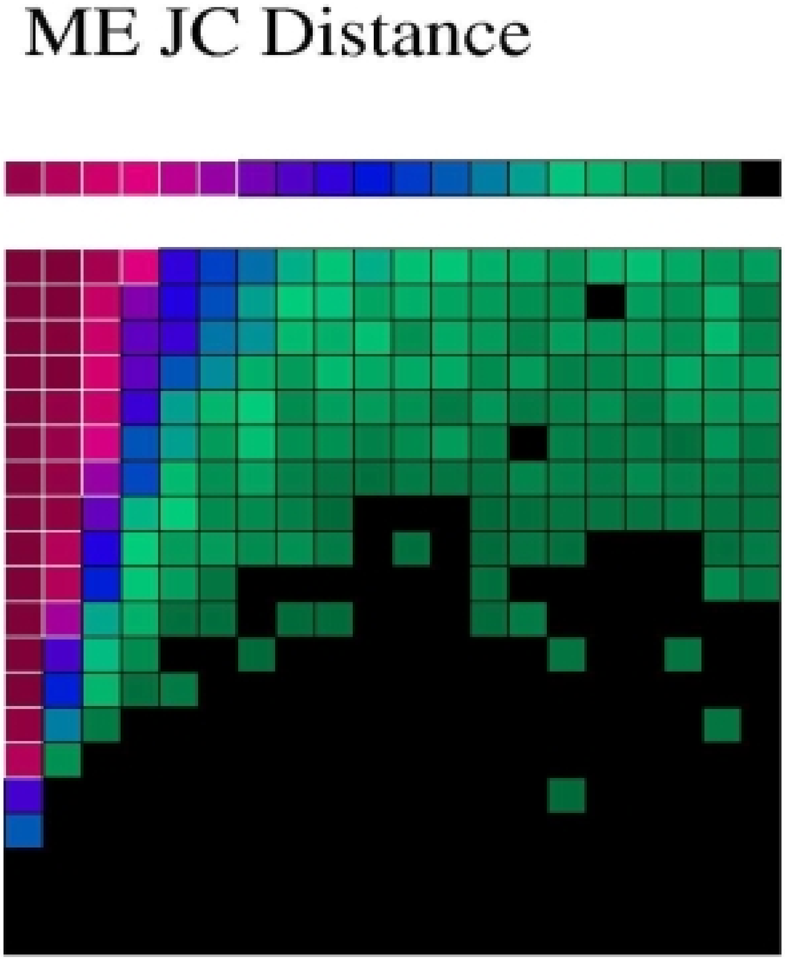


Figure 7.1

Figure 7.1 Performance of minimum evolution with a JC model correction of distances over the parameter space shown in Figure 3.2.

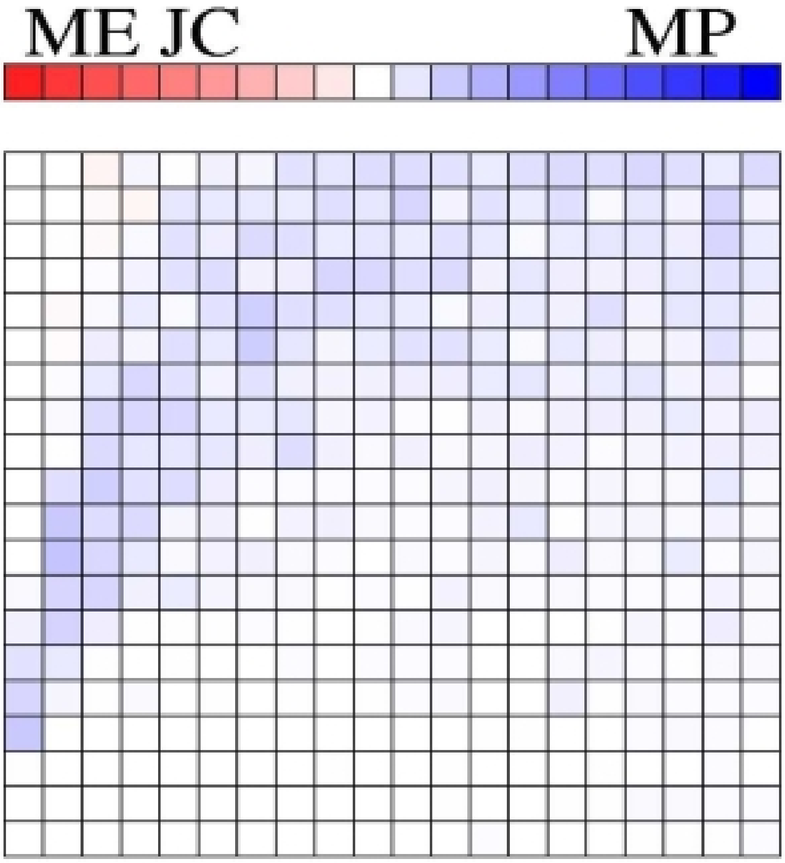


Figure 7.2

Contrast of minimum evolution with just a JC model correction to weighted parsimony made by subtracting the number of successes by minimum evolution from the number of successes by parsimony over the parameter space shown in Figure 3.2. Blue areas represent conditions in which parsimony is doing better than minimum evolution.

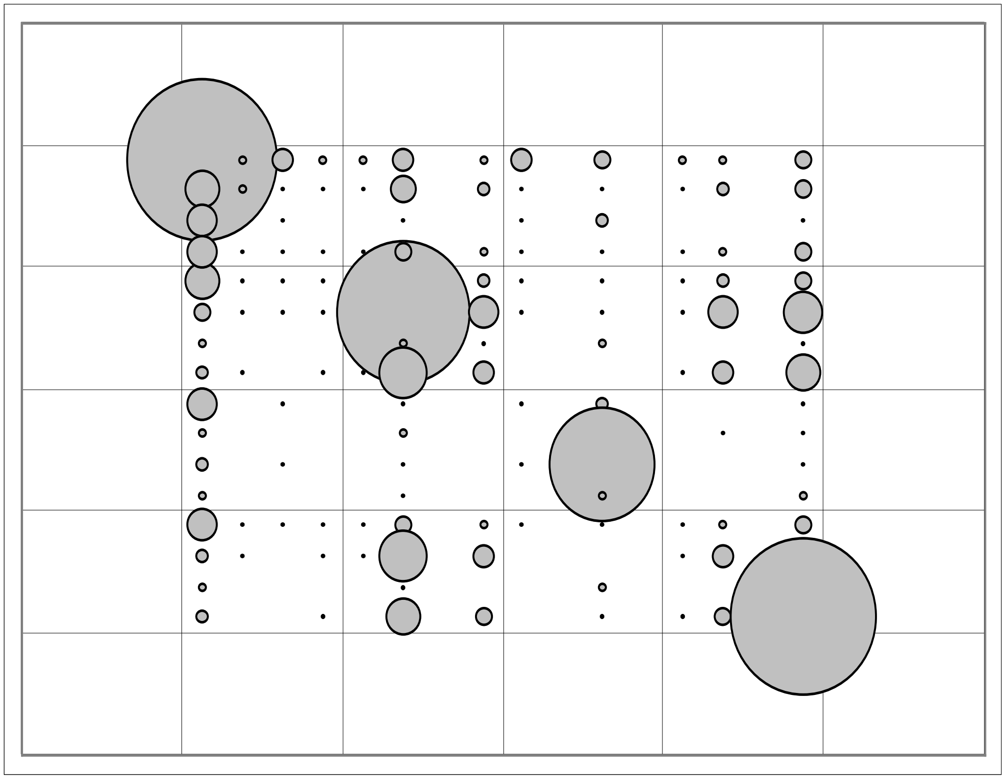


Figure 7.3

The spectrum of the HB model from the four taxa trees described in the text. The size of the circle is proportional the frequency of a particular data pattern. The true tree is ((1 , 2) , (3 , 4)), where 1 and 4 are the taxa on the ends of long branches. In the figure there are 16 columns and sixteen rows. The DNA bases are coded as A=0, C=1, G=2, and T=3. To determine the bases for a particular circle in the graphs count the columns starting with the left most column as zero until you reach the column with the desired circle in it. The column number modulo 4 (the remainder when you divide the column number by 4) codes for the base of taxon 2. The column number divided by 4 and rounded down codes for the base of taxon 3. Similarly the states of taxon 1 and 4 are determined by the row number (the top row is 0). The row modulo 4 is the state for taxon 1, and the row number divided by 4 and rounded down is the code for the state of taxon 4.

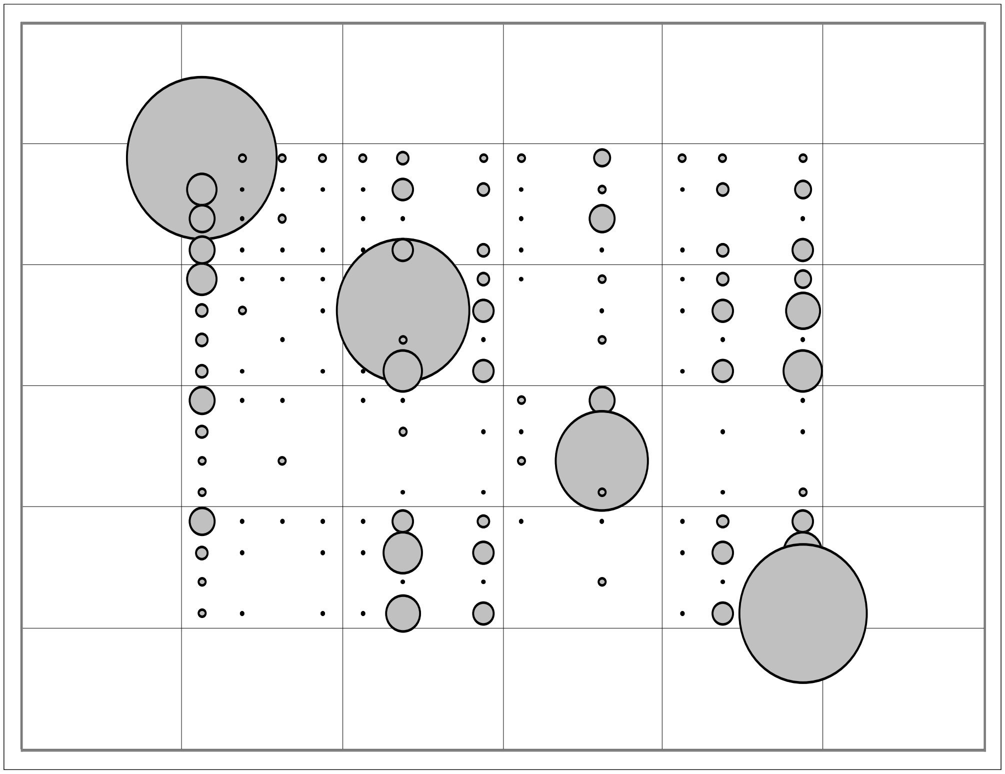


Figure 7.4

The spectrum of the GTR + G+ I model fit to an infinite sample of the spectrum shown in Figure 7.3. The size of the circle is proportional the frequency of a particular data pattern. The true tree is ((1 , 2) , (3 , 4)), where 1 and 4 are the taxa on the ends of long branches. In the figure there are 16 columns and sixteen rows. The DNA bases are coded as A=0, C=1, G=2, and T=3. To determine the bases for a particular circle in the graphs count the columns starting with the left most column as zero until you reach the column with the desired circle in it. The column number modulo 4 (the remainder when you divide the column number by 4) codes for the base of taxon 2. The column number divided by 4 and rounded down codes for the base of taxon 3. Similarly the states of taxon 1 and 4 are determined by the row number (the top row is 0). The row modulo 4 is the state for taxon 1, and the row number divided by 4 and rounded down is the code for the state of taxon 4.

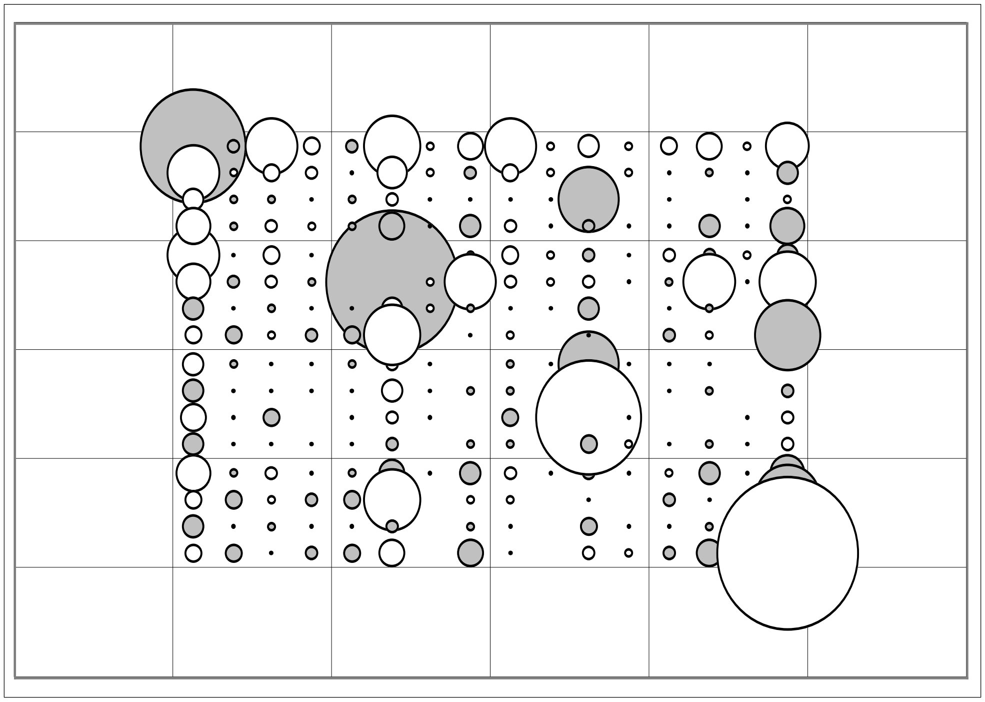


Figure 7.5

The difference between the two spectra shown in Figure 7.3 and 7.4 White circles indicate that GTR +G+I predicts too few of the pattern, Grey circles indicate the pattern is too common in the GTR+G+I spectrum.

References

Adachi, J and M Hasegawa. 1996. “Model of amino acid substitution in proteins encoded by mitochondrial DNA” J Mol Evol. 42(4):459-68.

Armstrong, LA, C Krajewski, and M Westerman. 1998. “Phylogeny of the dasyurid marsupial genus Antechinus based on cytochrome *b*, 12S rRNA, and protamine P1 genes” J Mammalogy. 79:1379-1389.

Bishop, MJ and AE Friday. 1985. “Evolutionary trees from nucleic acids and protein sequences” Proc R Soc London. B Biol Sci. 226:271-302.

Blacket, MJ andC Krajewski, A Labrinidis, B Cambron, S Cooper, and M Westerman. 1999. “Systematic relationships within the dasyurid marsupial tribe Sminthopsini--a multigene approach” Mol Phylogenet Evol. 12(2):140-55.

Brent, RP. 1973. *Algorithms for Minimization without Derivatives* (Englewood Cliffs, NJ: Prentice-Hall), Chapter 5.

Bruno, WJ and AL Halpern. 1999. “Topological bias and inconsistency of maximum likelihood using wrong models” Mol Biol Evol. 16(4):564-6.

Catzeflis, FM, C Hanni, P Sourrouille, and E Douzery. 1995. “Molecular systematics ofhystricognath rodents: the contribution of sciurognath mitochondrial 12S rRNA sequences” Mol Phylogenet Evol. 4(3):357-60.

Colgan DJ. 1999. “Phylogenetic studies of marsupials based on phosphoglyceratekinase DNA sequences” Mol Phylogenet Evol. 11(1):13-26.

Darwin, C. 1859. *On the Origin of Species by Means of Natural Selection, or the Preservation of Favoured Races in the Struggle for Life*.

Dayhoff, MO, RM Schwartz, and BC Orcutt. 1978. “A model of evolutionary change in proteins” in *Atlas of Protein Sequences and Structure* volume 5 suppl 3 (National Biomedical Research Foundation, Washington, D. C.) pp 345-352.

Dubois, J F, FM Catzeflis, and JJ Beintema. 1999. “The Phylogenetic Position of ‘Acomyinae’ (Rodentia, Mammalia) as Sister Group of a Murinae +

Gerbillinae Clade: Evidence from the Nuclear Ribonuclease Gene" Mol Phylogenet Evo. 13(1): 181-192.

Engel SR, KM Hogan, JF Taylor, and SK Davis. 1998. "Molecular systematics and paleobiogeography of the South American sigmodontine rodents" Mol Biol Evol. 15(1):35-49.

Felsenstein, J. 1978. "Cases in which parsimony or compatibility methods will be positively misleading" Systematic Zoology. 27: 401-410.

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

Felsenstein, J. 1981. "A likelihood approach to character weighting and what it tells us about parsimony and compatibility" Bio J Linnean Soc. 16: 183-196.

Flynn JJ, MA Nedbal, JW Dragoo, and RL Honeycutt. 2000. "Whence the red panda?" Mol. Phylogenet Evol. 17(2):190-199.

Galtier N. 2001. "Maximum-likelihood phylogenetic analysis under a covarion-like model" Mol Biol Evol. 18(5):866-73.

Galtier N, and M Gouy. 1998. "Inferring pattern and process: maximum-likelihood implementation of a nonhomogeneous model of DNA sequence evolution for phylogenetic analysis" Mol Biol Evol. 15(7):871-9.

Gaut, BS, and Lewis, P. O. 1995. "Success of maximum likelihoodphylogeny inference in the four taxon case" Mol Biol and Evol. 12: 152-62.

Goldman N, Thorne JL, and Jones DT. 1998. "Assessing the impact of secondarystructure and solvent accessibility on protein evolution" Genetics. 149(1):445-58.

Goldman N, Thorne JL, and Jones DT. 1996. "Using evolutionary trees in protein secondary structure prediction and other comparative sequence analyses" J Mol Biol. 263(2):196-208.

Goldman N and Yang Z. 1994. " A codon-based model of nucleotide substitution for protein-coding DNA sequences" Mol Biol Evol. 11(5):725-36.

Grantham R. 1974. “Amino acid difference formula to help explain protein evolution” Science. 185(154):862-4.

Graybeal, A. 1998. “Is it better to add taxa or characters to a difficult phylogenetic problem?”Systematic Biology, 47 (1):9-17.

Halpern AL and WJ Bruno. 1998. “Evolutionary distances for protein-coding sequences: modeling site-specific residue frequencies” Mol Biol Evol. 15(7):910-7.

Hasegawa M, H Kishino, and T Yano. 1985. “Dating of the human-ape splitting by a molecular clock of mitochondrial DNA” J Mol Evol. 22(2):160-74.

Hastad, O and M Bjoklund. 1998. “Nucleotide substitution model and estimation of phylogeny” Mol Biol Evo. 15:1381-1389.

Hendy, MD and D Penny. 1989. “A Framework for the Quantitive Study of Evolutionary Trees” Systematic Zoology, 38(4) 297-310.

Hendy, MD and D Penny. 1993. “Spectral analysis of phylogenetic data” J Class. 10:5-24.

Hillis DM. 1996. “Inferring complex phylogenies” Nature. 383(6596):130-1.

Hillis, DM. 1998. “Taxonomic sampling, phylogenetic accuracy, and investigator bias” Systematic Biology. 47(1):3-8.

Huelsenbeck, JP and DM Hillis. 1993. “Success of phylogenetic methods in the four taxon case” Systematic Biology. 42(3):247-265.

Huelsenbeck, JP, and M Kirkpatrick. 1996. “Do phylogenetic methods produce trees with biased shapes?” Evolution. 50(4):1418-1424.

Huelsenbeck, JP. 1995a. “The performance of phylogenetic methods in simulation” Systematic Biology. 44(2):17-48.

Huelsenbeck, JP. 1995b. “The robustness of two phylogenetic methods: Four taxon simulations reveal a slight superiority of maximum likelihood over neighbor joining” Mol Biol Evo. 12(5):843-849.

Huelsenbeck, JP, B Larget, and DL Swofford. 2000. “A compound Poisson process for relaxing the molecular clock” Genetics. 154(4):1879-1892.

Jones DT, WR Taylor, and JM Thornton. 1992. “The rapid generation of mutation data matrices from protein sequences” Comput Appl Biosci. 8(3):275-82.

Jukes, TH and CR Cantor. 1969. “Evolution of protein molecules.” in *Mammalian Protein Metabolism,* HN Munro (ed.) (Academic Press, New York) pp 21-132.

Kimura M. 1980. “A simple method for estimating evolutionary rates of base substitutions through comparative studies of nucleotide sequences.” J Mol Evol. 16(2):111-20.

Kishino H, and M Hasegawa. 1989. “Evaluation of the maximum likelihood estimate of the evolutionary tree topologies from DNA sequence data, and the branching order in hominoidea” J Mol Evol. 29(2):170-9.

Kishino H, JL Thorne, and WJ Bruno. 2001. “Performance of a divergence time estimation method under a probabilistic model of rate evolution” Mol Biol Evol. 18(3):352-61.

Lake JA. 1994. “Reconstructing evolutionary trees from DNA and protein sequences: paralinear distances” Proc Natl Acad Sci USA. 91(4):1455-9.

Lanave C, G Preparata, C Saccone, and G Serio. 1984. “A new method for calculating evolutionary substitution rates” J Mol Evol. 20(1):86-93.

Lara MC, Patton JL, and MN da Silva. 1996. “The simultaneous diversification of South American echimyid rodents (Hystricognathi) based on complete cytochrome *b* sequences” Mol Phylogenet Evol. 5(2):403-13.

LeDuc, RG, WF Perrin, and AE Dizon. 1999. “Phylogenetic relationships among delphinid cetaceans based on full cytochrome b sequences” Marine Mammal Science 15(3):619-648.

Lessa EP and JA Cook. 1998. “The molecular phylogenetics of tuco-tucos (genus Ctenomys, Rodentia: Octodontidae) suggests an early burst of speciation” Mol Phylogenet Evol. 9(1):88-99.

Lockhart, PJ, MA Steel, D Penny, and MD Hendy. 1994. “Recovering evolutionary trees under a more realistic model of sequence evolution.” Mol Biol Evol. 11(4): 605-612.

Madsen, O, M Scally, CJ Douady, DJ Kao, RW DeBry, R Adkins, HM Amrine, MJ Stanhope, WW de Jong, and MS Springer. 2001. “Parallel adaptive radiations in two major clades of placental mammals” Nature. 409(6820):610-4.

Matthee CA, and SK Davis.2001. “Molecular Insights into the Evolution of the Family Bovidae: A Nuclear DNA Perspective” Mol Biol Evol. 18(7):1220-30.

Miyata, T, S Miyazawa, and T Yasunaga. 1979. “Two types of amino acid substitutions in protein evolution” J Mol Evol. 12(3):219-36.

Murphy, WJ, E Eizirik, WE Johnson, YP Zhang, OA Ryder, and SJ O'Brien. 2001. “Molecular phylogenetics and the origins of placental mammals” Nature. 409(6820):614-8.

Muse SV. 1995. “Evolutionary analyses of DNA sequences subject to constraints of secondary structure” Genetics. 139(3):1429-39.

Muse SV, and BS Gaut. 1994. “A likelihood approach for comparing synonymousand nonsynonymous nucleotide substitution rates, with application to the chloroplast genome” Mol Biol Evol. 11(5):715-24.

Nedbal MA, MW Allard, and RL Honeycutt. 1994. “Molecular systematics of hystricognath rodents: evidence from the mitochondrial 12S rRNA gene” Mol Phylogenet Evol. 3(3):206-20.

Painter, J, C Krajewski,, and M. Westerman. 1995. “A molecular phylogeny of planigales” J. Mammal. 76: 406–413.

Poe, S. and DL Swofford. 1999. “Taxon sampling revisited” Nature. 398(6725): 299-300.

Powell, M. 1964. “An efficient method for finding the minimum of a function of several variables without calculating derivatives” Computer Journal. 7:155-62.

Rannala, B, J P Huelsenbeck, Z Yang, and R Nielsen. 1998. “Taxon sampling and the accuracy of large phylogenies” Systematic Biology. 47:702-709.

Steel, MA. 1994. “Recovering a tree from the leaf colourations it generates under a Markov model.” Applied Mathematics Letters, 7(2):19-24.

Steel, MA and D Penny. 2000. “Parsimony, likelihood and the role of models in molecular phylogenetics” Mol Biol Evo. 17(6):839-850.

Thorne JL, H Kishino, and IS Painter. 1998. “Estimating the rate of evolution of the rate of molecular evolution” Mol Biol Evo. 15(12):1647-57.

Thorne JL, N Goldman, and DT Jones. 1996. “Combining protein evolution and secondary structure” Mol Biol Evo. 13(5):666-73.

Tillier ER, and RA Collins. 1998. “High apparent rate of simultaneous compensatory base-pair substitutions in ribosomal RNA” Genetics. 148(4):1993-2002.

Tuffley, C and MA Steel. 1997. “Links between maximum likelihood and maximum parsimony under a simple model of site substitution.” Bull Math Biol. 59(3): 581-607.

Waddell, PJ and MA Steel. 1995. “General time reversible distances allowing a distribution of rates across sites” Research Report, Dept of Mathematics and Statistics, Canterbury University, 1995.

Waddell, PJ, N Okada, and M Hasegawa, 1999. “Toward resolving the intraordinal relationships of placental mammals” Systematic Biol. 48(1):1-5 .

Wilbur, WJ 1985. “On the PAM matrix model of protein evolution.” Mol Biol Evo. 2(5):434-47.

Yang Z. 1994. “Maximum likelihood phylogenetic estimation from DNA sequences with variable rates over sites: approximate methods” J Mol Evol. 39(3):306-14.

Yang Z. 1994. “Estimating the pattern of nucleotide substitution” J Mol Evol. 39(1):105-11.

Yang Z, N Goldman, and AE Friday. 1994. “Comparison of models for nucleotide substitution used in maximum-likelihood phylogenetic estimation” Mol Biol Evol. 11(2):316-24.

Yang, Z, N Goldman, and AE Friday. 1995. “Maximum likelihood trees from DNA sequences: a peculiar statistical estimation problem” Systematic Biology 44:384-399.

Yang, Z, R Nielsen, and M Hasegawa. 1998. “Models of amino acid substitution and applications to mitochondrial protein evolution” Mol Biol Evo 15:1600-1611.

Yang Z and AD Yoder. 1999. “Estimation of the transition/transversion rate bias and species sampling” J Mol Evol. 48(3):274-83.

Zhang, Z, L Huang, VM Shulmeister, YI Chi, KK Kim, LW Hung, AR Crofts, EA Berry, SH Kim. 1998. “Electron transfer by domain movement in cytochrome bc1” Nature. 392(6677):677-84.

VITA

Mark Travis Holder was born July 13, 1972 in Boston, Massachusetts, to Phyllis Holder and Leonard Donald Holder, Jr. He graduated from Nürnberg American High in 1990. Mark received a Bachelors of Science with majors in biochemistry and genetics from Texas A & M University at College Station, Texas in 1994, where he also worked as a technician in the lab of Dr. Scott Davis, and met his wife Kristina Kichler. In 1995 he entered graduate school at the University of Texas at Austin.

Permanent Address: 117 West Creek, Salado, TX 76571

This dissertation was typed by the author.