CONSTRUCTION OF AMINO ACID RATE MATRICES AND EXTENSIONS OF THE BARRY AND HARTIGAN MODEL FOR PHYLOGENETIC INFERENCE

by

Liwen Zou

Submitted in partial fulfillment of the requirements
for the degree of Doctor of Philosophy

at

Dalhousie University
Halifax, Nova Scotia
August 2011

DALHOUSIE UNIVERSITY

DEPARTMENT OF MATHEMATICS AND STATISTICS

The undersigned hereby certify that they have read and recommend to the Faculty of Graduate Studies for acceptance a thesis entitled "CONSTRUCTION OF AMINO ACID RATE MATRICES AND EXTENSIONS OF THE BARRY AND HARTIGAN MODEL FOR PHYLOGENETIC INFERENCE" by Liwen Zou in partial fulfillment of the requirements for the degree of Doctor of Philosophy.

Dated: August 9, 2011

External Examiner: ______________________________

Research Supervisor: ______________________________

Examining Committee: ______________________________

Departmental Representive: ______________________________

DALHOUSIE UNIVERSITY

DATE: August 9, 2011

AUTHOR: Liwen Zou

TITLE: CONSTRUCTION OF AMINO ACID RATE MATRICES AND EXTENSIONS OF THE BARRY AND HARTIGAN MODEL FOR PHYLOGENETIC INFERENCE

DEPARTMENT OR SCHOOL: Department of Mathematics and Statistics

DEGREE: PhD CONVOCATION: October YEAR: 2011

Signature of Author

For my family Tieshan, Xuetian and Xiaoqun

Table of Contents

List of Tables

List of Figures

Abstract

This thesis considers two distinct topics in phylogenetic analysis. The first is construction of empirical rate matrices for amino acid models. The second topic, which constitutes the majority of the thesis, involves analysis of and extensions to the BH model of Barry and Hartigan (1987).

There are a number of rate matrices used for phylogenetic analysis including the PAM (Dayhoff et al. 1979), JTT (Jones et al. 1992) and WAG (Whelan and Goldman 2001). The construction of each of these has difficulties. To avoid adjusting for multiple substitutions, the PAM and JTT matrices were constructed using only a subset of the data consisting of closely related species. The WAG model used an incomplete maximum likelihood estimation to reduce computational cost. We develop a modification of the pairwise methods first described in Arvestad and Bruno that better adjusts for some of the sparseness difficulties that arise with amino acid data.

The BH model is very flexible, allowing separate discrete-time Markov processes to occur along different edges. We show, however, that an identifiability problem arises for the BH model making it difficult to estimate character state frequencies at internal nodes. To obtain such frequencies and edge-lengths for BH model fits, we define a nonstationary GTR (NSGTR) model along an edge, and find the NSGTR model that best approximates the fitted BH model. The NSGTR model is slightly more restrictive but allows for estimation of internal node frequencies and interpretable edge lengths.

While adjusting for rates-across-sites variation is now common practice in phylogenetic analyses, it is widely recognized that in reality evolutionary processes can change over both sites and lineages. As an adjustment for this, we introduce a BH mixture model that not only allows completely different models along edges of a topology, but also allows for different site classes whose evolutionary dynamics can take any form.

List of Abbreviations and Symbols Used

Abbreviation	**Description**
A:	Adenine.
AB:	The method in Arvestad and Bruno (1997).
aveBias:	The averaged bias.
aveMSE:	The averaged MSE.
BH:	Nonstationary model in Barry and Hartigan (1987).
BHK:	K-class mixture BH model.
BH(K-1):	A mixture BH model which has K-1 BH classes and one invariable class.
BRKALN:	D. Jones, unpublished database.
C:	Cytosine.
DCFreq:	An instantaneous rate matrix which is directly calculated from observed frequencies Kosiol and Goldman (2005).
DCMut:	An instantaneous rate matrix which is calculated using the mutabilities in Dayhoff et al. (1979).
DNA:	Deoxyribonuleic acid.
EM:	Expectation and maximization.
G:	Guanine.
GG:	Nonstationary model in Galtier and Gouy (1995).
GTR:	General time-reversible Markov model.
GTR+Γ+I:	Mixture Markov model in which all edges in one class share a GTR model and the sites have a Γ distribution and the invariable sites are considered separately.
HKY85:	Time-reversible Markov model in Hasegawa et al. (1985).
JTT:	The method in Jones et al. (1992).
LogDet:	A measure for evolutionary distance.
MSE:	The mean squared errors.
ML:	Maximum likelihood.
nhPhyML:	A free phylogenetic analysis software which implemented the GG model in Boussau and Gouy (2006).
NSGTR:	Nonstationary GTR model.
PANDIT:	A protein sequence database of Whelan et al. (2006).
PAM:	The method in Dayhoff et al. (1979).
PAML:	A free phylogenetic analysis package in Yang (2007).
PAUP*:	A commercial software package for Phylogenetic analysis in Wilgenbusch (2003).
PHYLIP:	A free phylogenetic analysis package in Felsenstein (1989).
PhyML:	A free phylogenetic analysis package in Guindon and Gascuel (2003).
SWISS-PROT:	A protein knowledgebase which is manually annotated and reviewed in Boeckmann et al. (2003).

Abbreviation	**Description**
T92:	The time-reversible model in Tamura in (1992).
TREE-PUZZLE:	A free phylogenetic analysis package in Schmidt et al. 2002.
T:	Thymine.
WAG:	The method in Whelan and Goldman (2001).
YR:	The nonstationary model in Yang and Roberts (1995).

Acknowledgements

First all, I would like to thank my supervisors Dr.Chris Field, Dr.Andew Roger, and Dr.Edward Susko for leading me into the phylogenetic research area. They have very generously contributed their time and been very patient over this long and happy journey. I enjoyed every minutes over this long way. Thanks for all of their very creative and thoughtful comments and suggestions. Those are all good guides and provided tremendous help. Their enthusiasm and faithfulness to research are always the inspiration and the encouragement. Without their significant constributions, this thesis might never have been finished.

Special thanks to Dr.Joseph Bielawski and Dr.Bruce Smith. Dr.Bielawski has read entire thesis very carfully and given very constructive comments and suggestions. Dr.Bruce Smiths encouragement has been a great help during this long journey.

Thanks to Dr.Vivek Jayswal for many discussions and suggestions. Thanks to Dr.Elisabeth Tillier, Dr.Robert Beiko and Dr.Christian Blouin for their useful comments and suggestions.

My family were always behind me and gave tremedous support. Their patience has been a tremendous help especially during difficult times.

My thanks and appologies go to everyone who has helped me that I have not named here.

Chapter 1

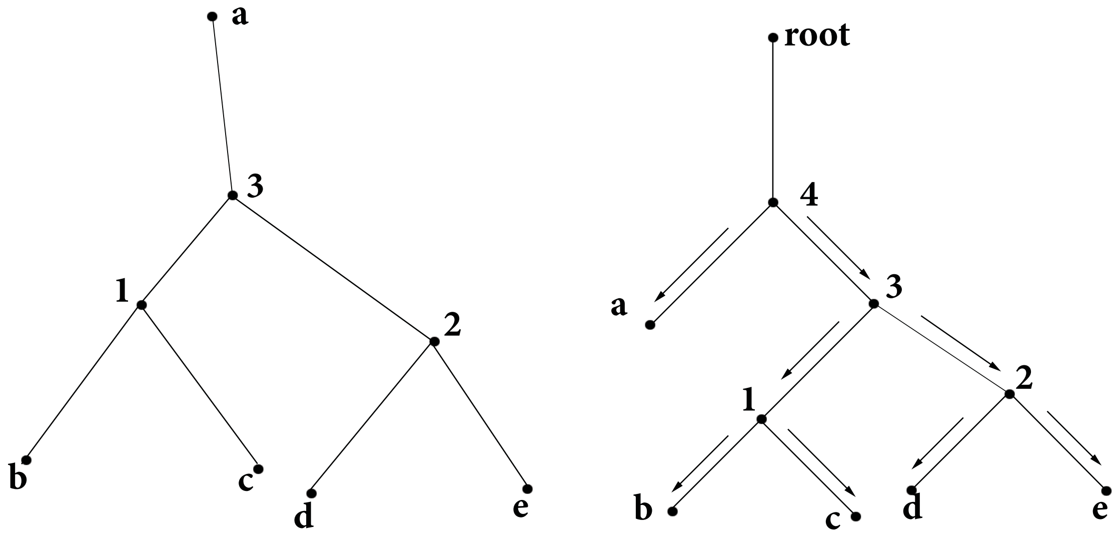
Introduction

Exploring how evolution has proceeded at the fundamental molecular level is the ultimate goal of this work. Currently, the events involved in the origin of life remain mysterious as do many of the detailed mechanisms by which diverse organisms have since evolved over the past 3.5 billion years of Earth's history. In this thesis, methods for addressing these questions are presented that include the construction and fitting of more realistic models of the molecular evolutionary processes.

1.1 Reconstructing Evolutionary History Usingc Molecular Data

1.1.1 The Phylogenetic Tree

A phylogenetic tree is composed of nodes and edges that depict the inferred evolutionary relationships among species and their ancestors. The accurate estimation of such trees from DNA and protein sequence data is a main goal of this thesis. An example phylogeny is given in Figure 1.1 that depicts both rooted and unrooted versions of the same tree. Nodes a, b, c, d, e in both trees in Figure 1.1 are the leaves or external nodes which represent the observed taxa for which we have protein or DNA sequence data; nodes 1, 2 and 3 in Figure 1.1(a) and nodes 1, 2, 3 and 4 in Figure 1.1(b) are internal nodes which correspond to ancestral taxa and therefore have no available data. With m taxa, there are $m-2$ internal nodes and $2m-3$ edges in an unrooted tree whereas there are $m-1$ internal nodes and $2m-2$ edges in a rooted tree. In a rooted tree as in Figure 1.1(b), the direction of evolution is specified, while the evolutionary direction of an unrooted tree is not specified.



a

b

Figure 1.1: Illustration of rooted and unrooted trees. The graph in (a) is an unrooted version of the rooted tree in (b)

1.1.2 Sequence Data

The study of molecular evolution focuses on the genetic material, DNA, which carries the hereditary information of all living things. The specific DNA sequences of interest are called genes which determine the characteristics of an organism. The hereditary information in genes can pass from generation to generation. During evolution, the nucleotides in a gene may be mutated into the other nucleotides, deleted or inserted. Nucleotides are distinguished by which of the four chemical bases they contain: adenine (A) or guanine (G) in the purine group, and thymine (T) or cytosine (C) in the pyrimidine group. Phylogenetic analysis of DNA sequence data can be performed on one or multiple genes. Protein sequences, in contrast, are made up of sequences of amino acids that are typically inferred from protein-coding DNA sequences via the genetic code. This triplet coding scheme has three consecutive nucleotides that collectively represent one codon. Each of the 61 of the 64 unique three-nucleotide codons specify at most one of 20 amino acids with the remaining three representing 'stop' signals at the end of genes signaling the termination of protein synthesis (Chapter 1 in Graur and Li 2000).

Because of lineage-specific insertions and deletions, the i^{th} position or site of two sequences need not be evolutionarily related. Before a phylogenetic tree can be inferred from molecular sequences from multiple species in a data set, the sequences must be aligned to that the i^{th} site refers to the same nucleotide position of the sequence for the common ancestor of all of the observed sequences. Stretches of nucleotides that have a common pattern across all species often arise through constraints due to similar or identical biochemical functions and thus aid in the alignment process.

In sequence analysis, similar regions in the sequences may indicate a similar or identical biochemical function. Table 1.1 is an example of a DNA sequence alignment. When aligning six sequences in Table 1.1, these six sequences are assumed to have the same ancestor. A gap, shown as "-" such as at sites 9, 10, 11 of 'species5' in Table 1.1, may be inserted when aligning sequences so that more nucleotides are shared in common across species (Mount 2004). Any gap in an alignment is typically interpreted as 'missing data' that corresponds to an unknown nucleotide (or amino acid). Normally, alignments are obtained either manually or by using various computer programs such as Chenna et al. (2003). Although the generation of alignments may introduce errors regardless of which methods are used, it is usually assumed that such errors are either minor, or that error-prone regions of alignments can be recognized and removed prior to phylogenetic analysis.

Table 1.1: An example of an alignment

								sites							
species	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15
species1	C	A	A	T	G	A	A	C	A	G	A	A	G	T	T
species2	C	A	A	T	G	A	A	C	A	G	A	A	A	T	T
species3	C	A	A	T	G	A	A	C	A	G	A	A	A	T	T
species4	C	A	A	T	G	A	A	C	A	G	A	A	A	T	T
species5	C	A	A	T	G	A	A	C	-	-	-	C	A	T	T
species6	C	A	A	T	G	A	A	C	A	G	A	C	A	T	T

Once an alignment is obtained, the process of modeling evolutionary processes over a phylogeny can begin. To simplify computation and subsequent analysis, sequence alignment columns (here referred to as the sites) such as those numbered $1, \ldots 15$ in Table 1.1 are considered to be independent and identically distributed (i.i.d). The character vector observed at each site, e.g., 'CCCCCC' in site 1, is called a ‘site pattern’. A site is a variable site if the character states change over evolutionary time whereas a site is invariable if it cannot change over time. Invariable sites are often the result of an essential function performed by the nucleotide or amino acid at that particular position in the molecule under consideration. Note that sites in which the same character state is observed for all species are denoted as invariant sites, e.g., sites 1, 2, 3, and etc; the latter are either invariable (i.e., they cannot change) or a variable site that simply was not observed to change for the sample of taxa under investigation.

1.1.3 Modeling Evolutionary Processes

In this thesis the evolutionary processes along the edges of a phylogenetic tree, e.g., the process from node a to node 3 in Figure 1.1(a), are of particular interest. When modeling evolution, it is usually reasonable to assume that the processes from any ancestor to its two descendants are independent of each other. For instance, the process from node 1 to node b (Figure 1.1(b)) is independent of the process operating from node 1 to node c. Furthermore, the state at node b is only dependent on the state of its ancestral node 1, but not other nodes on the evolutionary path from the root to node b such as nodes 3, 4 and the root.

If one assumes a stationary model that does not vary over the tree and that two processes along opposite directions on every edge, probabilistically, are the same, the location of the root of the tree becomes unimportant in terms of the likelihood calculations described below. In other words, differently rooted trees under this assumption will yield the same likelihoods as the corresponding unrooted tree. If the two processes in opposite directions along an edge are not the same, the location of the root matters and a nonstationary model is appropriate.

Another important consideration in this thesis is the presence of distinct evolutionary processes operating at different sites in data set. Models for the evolutionary processes among sites often assume distinct overall rates of evolution occur at different sites (rates-across-sites) with the rates at sites treated as unknown and modeled by a variable rates distribution.

In this thesis, we focus on nonstationary and rates-across-sites variation models when analyzing DNA and amino acid data. Since the stationary model is an important concept and will be used in nonstationary models when describing processes along edges, we start by discussing the stationary models and then introducing nonstationary models.

1.2 Stationary Models

1.2.1 Stationary Models for DNA Data

The stationary models used for modeling DNA data are time-reversible Markov models with the transition probability matrix $P(t) = e^{Rt}$ between two taxa separated by an evolutionary distance of t. The matrix R is the instantaneous rate matrix. Its ij^{th} off-diagonal entry is the substitution rate of replacing state i by state j and R has the property $\sum_j r_{ij} = 0$. In a transition matrix $P(t)$, the ij^{th} entry is the probability that given ancestral state i, the descendant has the state j after t evolutionary time units later, and row sums are equal to 1. Time-reversibility implies that the entries in the instantaneous rate matrix R satisfies $r_{ij}\pi_i = r_{ji}\pi_j$, where π_i are the stationary frequencies. It is often convenient to decompose R as $R = A\Pi$, where A is a symmetric matrix, referred to as the exchangeability matrix, and Π is a diagonal matrix with the diagonal elements π_i. In a stationary model, the frequency vectors at all nodes are the same as the stationary frequency vector; the edge lengths can be interpreted as the expected numbers of substitutions along edges if the estimate of the rate matrix satisfies $-\sum_j \pi_j R_{jj} = 1$, which means that the average of the rates in R is one (cf. Chapter 13 Felsenstein 2004). The main differences between the various stationary Markov models used in practice are due to differing constraints on the state exchangeabilities and stationary frequencies.

The general time reversible Markov model (Tavaré, 1986), referred to as the GTR model, is the most flexible model amongst stationary models since it has no additional constraints on the exchangeabilities and stationary frequencies other than those articulated above. For DNA sequence data, there are 9 parameters in a GTR model, 3 for stationary frequencies and 6 for exchangeabilities.

Two examples of time-reversible models with additional constraints that we will consider later are the model in Hasegawa et al. (1985), referred to as the HKY85 model, and the model in Tamura (1992), referred to as the T92 model. The HKY85 model has common rates for the transition process and for the transversion process. The substitutions between two states of purine group $A \leftrightarrow G$ or between two states of pyrimidine group $C \leftrightarrow T$ are called transitions; the substitutions between states of purine and pyrimidine groups $A \leftrightarrow C$, $A \leftrightarrow T$, $G \leftrightarrow C$ and , $G \leftrightarrow T$ are called transversions. The exchangeability rates of transitions within purine and pyrimidine groups are assumed to be the same, $a_{AG} = a_{CT}$ and the exchangeability rates for transitions between purine and pyrimidine groups are assumed to be equal, $a_{AC} = a_{AT} = a_{CG} = a_{GT}$. Under HKY85 there are no constraints on the stationary frequencies. In contrast, while the T92 model treats transitions and transversions in similar manner to HKY85, it makes the further assumption that the frequencies of character bases C and G are equal ($\pi_C = \pi_G$) and similarly that A and T are equal ($\pi_A = \pi_T$). Thus the T92 model is a special case of the HKY85 model.

The models in Yang and Roberts (1995) and Galtier and Gouy (1995, 1998) used stationary models to model the nonstationary processes along edges. For instance, consider the tree in Figure 1.1(b). When modeling the processes from root to leaf nodes, a stationary model for the process from ancestor to descendant applies for each edge. The root frequencies are specified which may or may not be the same as the stationary frequencies in the stationary model along the edge connecting the root and the internal node. In Yang and Roberts (1995), the HKY85 model was used as the stationary model along each edge. For each edge l, a stationary frequency vector π^l was estimated; for all edges, a transition and transversion rate was estimated and the root node frequency vector was also estimated. In Galtier and Gouy (1995, 1998), the T92 model was used as the stationary model along edges in the same way as was done with the HKY85 model in Yang and Roberts (1995). In subsequent chapters, we will define a nonstationary version of the GTR model, referred to as the NSGTR model, that will fit different models for each edge.

1.2.2 Maximum Likelihood Estimation

We used maximum likelihood (ML) estimation for almost all parameter estimation in this thesis. Denote the sequence data $\underline{X} = \{\underline{x}_1 \dots \underline{x}_n\}$, where $\underline{x}_i$ is the data pattern at site i; n is the number of sites; m is the number of taxa. Let Θ be the parameters in the Markov model and let $T = \{t_1 \dots t_{2m-3}\}$ be the vector of edge lengths for a unrooted tree, where $t_1 \dots t_{2m-3}$ are the lengths of $2m - 3$ edges in an unrooted tree. The likelihood is

$$P(\underline{X}|\Theta, T) = \prod_{i=1}^{n} P(\underline{x}_i|\Theta, T) \tag{1.1}$$

Here $P(\underline{x}_i|\Theta, T)$ is sometimes referred to as the site likelihood. In ML estimation, all or some of the parameters Θ and T are estimated by maximizing $P(\underline{X}|\Theta, T)$.

Using ML estimation in phylogenetic analysis, we have to face two challenges: calculating likelihoods and optimizing the parameters. For each site pattern $\underline{x}_i$, because models of evolution are usually expressed as a sequence of conditionally independent models given ancestral data, the only probability that can be directly calculated is the probability $P(\underline{x}_i, x_1, \dots, x_{m-2}|\Theta, T)$, of both $\underline{x}_i$ and the ancestral data $x_1, \dots, x_{m-2}$ for $m - 2$ ancestors in an unrooted tree. The likelihood contribution for the site is obtained by summing over the unobserved ancestral states, $P(\underline{x}_i|\Theta, T) = \sum_{x_1 \dots x_{m-2} \in \mathcal{B}} P(\underline{x}_i, x_1, \dots, x_{m-2}|\Theta, T)$, where $\mathcal{B} = \{A, C, G, T\}$. In this sum, there are 4^{m-2} terms making direct combinations and, as a result, direct likelihood calculation for a tree is infeasible when the number of taxa m is large. The pruning algorithm (cf. Chapter 16 in Felsenstein 2004) is used to overcome this shortcoming. The basic idea of the pruning algorithm is to use the nesting rule and post-order traversal along a tree. For instance, for the tree in Figure 1.1a, the likelihood for a site pattern $\underline{x}$ given the parameters Θ and T is

$$\begin{aligned} P(\underline{x}|\Theta, T) &= \sum_{x_1,x_2,x_3\in\mathcal{B}} \pi_{x_a} P(x_3|x_a,\Theta,t_{a3})P(x_1|x_3,\Theta,t_{31})P(x_2|x_3,\Theta,t_{32}) \\ &\qquad P(x_b|x_1,\Theta,t_{1b})P(x_c|x_1,\Theta,t_{1c}) \\ &\qquad P(x_d|x_2,\Theta,t_{2d})P(x_e|x_2,\Theta,t_{2e}) \end{aligned}$$

where $t_{a3}, t_{31}, t_{32}, t_{1b}, t_{1c}, t_{2d}, t_{2e}$ are the lengths of edges $(a,3)$, $(3,1)$, $(3,2)$, $(1,b)$, $(1,c)$, $(2,d)$, $(2,e)$; x_i for $i = \{1,2,3,a,\dots e\}$ are the nucleotides at nodes. Using the pruning algorithm, $P(x|\Theta,T)$ can be calculated by

$$\begin{aligned} P(\underline{x}|\Theta, T) &= \pi_{x_a} \sum_{x_3\in\mathcal{B}} P(x_3|x_a,\Theta,t_{a3}) \\ &\qquad \left[\sum_{x_1\in\mathcal{B}} P(x_1|x_3,\Theta,t_{31})P(x_b|x_1,\Theta,T)P(x_c|x_1,\Theta,T)\right] \\ &\qquad \left[\sum_{x_2\in\mathcal{B}} P(x_2|x_3,\Theta,t_{32})P(x_d|x_2,\Theta,T)P(x_e|x_2,\Theta,T)\right] \end{aligned}$$

The calculation order in the above formula is to calculate the terms inside the two square brackets first, and then compute the summations over x_3 for a clear saving on computation cost. For example, for this particular five-taxon tree, using the pruning algorithm, the 32 terms must be calculated and summed over compared with the 64 terms calculated without the pruning algorithm. Since this algorithm visits children such as nodes b and c before visiting their parent node 1, it is referred to as post-order traversal.

Ideally, for tree estimation, the ML estimates of free parameters (e.g., exchangeabilities, edge lengths, etc.) are obtained for each possible topology. Difficulties arise from the fact that the tree topology is a discrete variable while edge lengths, exchangeabilities and frequencies are continuous variables. Sophisticated optimization procedures combining searches through very large discrete and continuous parameter spaces are therefore required. To make the computation feasible, heuristic methods are usually employed in practice that are not guaranteed to yield globally optimal estimates. ML estimation procedures for common models are implemented in software packages such as PAML (Yang, 2007), PAUP* (Wilgenbusch 2003), PHYLIP (Felsenstein 1989) and PhyML (Guindon and Gascuel 2003).

1.2.3 Empirical Rate Matrices for Modeling Amino Acid Data

Estimation of parameters of the substitution model (i.e., rate matrix parameters) becomes complicated with amino acid data. The increase from four character states to 20 character states for amino acid data potentially provides more information for the inference, however this comes at a cost due to the increase in model complexity. Likelihood computations become more intensive for amino acid data as the pruning algorithm now requires repeated summation over 20 rather than four terms. Furthermore, for an amino acid GTR model, there are 209 parameters in the exchangeability matrix and stationary frequencies compared to 9 for DNA data and ML estimation in that much larger parameter space is also more challenging. Consequently, a pre-determined fixed rate matrix is usually used for analyzing amino acid data to reduce the complexity of optimization during phylogeny estimation. Since a new method for the estimation of fixed empirical rate matrices will be the focus of Chapter 2, we review three commonly used models, the accepted point mutation (PAM) (Dayhoff et al. 1979), Jones-Taylor-Thornton (JTT) (Jones et al. 1992) and the Whelan-and-Goldman (WAG) (Whelan and Goldman 2001) matrices, in some detail here. These three models were obtained using protein families in databases with a wide coverage of protein families.

The PAM and JTT models were obtained by counting the substitutions using trees created with the parsimony method. As stated in Edwards and Cavalli-Sforza (1963), the parsimony evolution tree is constructed with the goal of minimizing the "net amount of evolution". When creating a phylogenetic tree, the parsimony method treats an observed difference in nucleotides as a consequence of a single substitution. The possibility of multiple substitutions is not adjusted for. Therefore it only works well in the sequences which has small edge lengths but not necessarily for large ones. For both PAM and JTT models, sequences with 85% or higher similarity were used to avoid the observed differences that do not correspond to single changes. The PAM model was obtained from 71 groups of closely related proteins; the JTT model was obtained using 559,692 pairs filtered from 130,000,000 pairs obtained from the SWISS-PROT database of Boeckmann et al. (2003).

Since the original elaboration of the PAM model did not define an instantaneous rate matrix, Kishino et al. (1990) and Kosiol and Goldman (2005) provided methods to obtain the rate matrices from this model. Kishino et al. (1990) provided an eigen decomposition solution: for a continuous-time model, the transition matrix

$P(t)$ has the eigen-decomposition, $Ue^{\Lambda t}U^{-1}$. The eigenvector matrix U and eigenvalue matrix Λ of $P(t)$ are the same as the ones of a corresponding instantaneous rate matrix. Therefore, from the eigen decomposition of a PAM matrix, the estimate of U is obtained directly by taking the eigenvectors of the PAM matrix; the estimate of Λ is obtained by normalizing the logarithms of eigenvalues of the PAM matrix so that the resulting rate matrix, $R = U\Lambda U^{-1}$, satisfies $-\sum_i \pi_i R_{ii} = 1$. In contrast, in Kosiol and Goldman (2005) two methods were proposed for calculating the rate matrix directly from a PAM matrix without any eigen decomposition requirement. The first method is called DCMut, Direct Computation with Mutabilities, that calculates the rate r_{ij}, replacing state i by state j, only using the mutabilities in the PAM model (Dayhoff et al. 1979). The second method these authors introduced is called DCFreq, Direct Computation with Frequencies, that calculates the rate r_{ij} using only the observed changes. Further details of these methods are given in Kosiol and Goldman (2005). The concern about the counting method is that all differences observed resulted from only one amino acid substitution. Wilbur (1985) showed that 85 percent similarity criterion does not guarantee that all observed differences were a consequence of a single amino acid substitution.

Müller and Vingron (2000) and Müller et al. (2002) provide a way to estimate a rate matrix and stationary frequency vector for divergent data. Under the GTR model, there are 209 parameters that require optimization for ML estimation. To avoid directly estimating these parameters, Müller et al. use an alternative characterization of a rate matrices via the resolvents discussed in Fukushima (1980).

Every rate matrix Q can be expressed as $Q = \alpha I - R_\alpha^{-1}$, for some α, where

$$R_\alpha = \int_0^\infty e^{-\alpha t} P(t) dt \tag{1.2}$$

is referred to as the resolvent associated with Q. Müller et el. use the fact that the resolvent is expressed as a function of the transition matrix $P(t)$ to derive a method for estimating it. Pairs of sequences with similar evolutionary distances are placed in bins and, for the kth bin, having an estimated representative distances, $\hat{t}_k$, the substitution matrix is estimated from the pairwise frequencies of amino acids for pairs in the bin. Linear interpolation between estimated $P(t)$ for the $\hat{t}_k$ is used to obtain $P(t)$ for all t, which is then substituted into (1.2) to obtain the resolvent for any given value of α. The remaining parameter, α, is obtained by maximizing the average likelihood, averaged over all pairs. Since the resolvent is obtained using piecewise linear approximation and only α is optimized, the approach is efficient by comparison with full GTR optimization. Since the pairwise distances are estimated using an initial rate matrix, the above procedure was iterated. In each iteration, the newly estimated rate matrix replaces the rate matrix used in previous iteration; iterations continued until no further changes to the rate matrix occurred. The methods were used to obtain an empirical rate matrix from 2.7 million pairs of aligned sequences in the SYSTERS database of Krause and Vingron (1998). A total of 80 bins were used that covered PAM distances from 1 to approximately 300.

Veerassamy et al. (2003) develop methods, referred to as PMB (Probability Matrix from Blocks), to estimate empirical rate matrices from estimates of joint probability matrices at differing evolutionary distances. The joint probability matrices they use come from the BLOSUM series of Henikoff and Henikoff (2002). These are converted to substitution matrices, M, and initial observed distances, sometimes referred to as p-distances, $1 - \sum_i \pi_i M_{ii}$, are obtained. Since there are multiple BLOSUM matrices, corresponding to different evolutionary distances, multiple p-distances can be obtained. Veerassamy et al. (2003) use an approximation to a differential equation relating usual evolutionary distance (expected numbers of substitutions) to p-distances. The equation is solved and the observed p-distance is substituted to obtain an estimate, $\hat{P}_c$, of the evolutionary distance for the cth BLOSUM substitution matrix, M_c. An estimate, A_c of the rate matrix is obtained through the matrix logarithm $\log(M_c)/\hat{P}_c$. Since this gives different rate matrices for different BLOSUM classes, a weighted average, $U = \sum w_c A_c$, is used to obtain a final estimate of the rate matrix. The weights are chosen to make small a sort of average distance between M_c and $e^{U\hat{P}_c}$, both of which are estimating the same quantity.

Likelihood methods for sequence data was first introduced in Neyman (1971). Since then it has come to be one of the most frequent used methods in phylogenetic analysis. For empirical rate matrix estimation, Adachi and Hasegawa (1996), Yang et al. (1998), and Adachi et al. (2000) constructed rate matrices from small data sets. Whelan and Goldman (2001) proposed a model (WAG) obtained from the unpublished BRKALN database in which there are 3905 sequences in 182 protein families. Because of the high computational cost, the WAG model didn't optimize all parameters using ML estimation. Prior to using ML obtaining the rate matrix, the topology of each protein family was estimated by a neighbor-joining method in Saitou and Nei (1987) using the pairwise distances obtained under the PAM model. Then the edge lengths were estimated using ML under the JTT model with the fixed topology obtained from the previous step for each protein family. Finally the instantaneous rates and stationary frequencies were obtained using ML while the topologies of protein families were fixed and the edge lengths for each protein family were fixed up to a single multiplicative factor; these multiplicative factors were jointly optimized with rates and frequencies.

The instantaneous rate model for amino acids in Le and Gascuel (2008), which we refer to as the LG model, was estimated through ML allowing for rates-across-sites variation. The data used for estimation were from the May 2007 version of Bateman et al. (2004) but restricting attention to protein families that have at least 5 taxa and 50 sites. There were 3,912 protein families that met those criteria. An approximate ML method was used instead of a full ML estimation. In their approach, estimation has three steps: (1) estimating topologies for protein families using ML estimation under the WAG+Γ 4+I model; (2) selecting the rate category which gives the largest likelihood among all rate categories for each site in each protein family; (3) estimating the rate matrix using a likelihood that treats the rate category at each site as fixed at the value determined in (2). This solution avoids the computation coming from summing over different rate categories as is required for a full likelihood calculation.

1.2.4 The Arvestad and Bruno Method

Arvestad and Bruno (1997) proposed a method (herein referred to as the AB method) to construct the instantaneous rate matrix under a time-reversible Markov chain framework for modeling DNA sequence evolution. Among the methods used for empirical rate matrix estimation, the new methods proposed in Chapter 2 are conceptually most similar to the AB method. The advantage of this type of approach is that no evolutionary tree construction is required for the estimation of the rate matrix.

The AB method estimates the rate matrix from K pairs of sequences. If one assumes that the rate matrix is the same for each pair, the eigen-decompositions of the transition matrix of the k^{th} pair, $P(t^k) = Ue^{\Lambda t^k}U^{-1}$ will give the eigenvector matrix U and the eigenvalues matrix Λ of the rate matrix R. Here, Λ is the diagonal eigenvalue matrix of the rate matrix with the diagonal elements λ_j and t^k is the evolutionary distance separating two taxa of the k^{th} pair. Thus the average $P(t^k)$ will give eigenvector matrix U. The AB method uses the averaged transition matrices obtained from data to get an estimate of U. With this estimate, for each pair, an estimate of $\lambda_j t^k$ can be obtained from the observed transition matrix for the k^{th} pair. While the estimates of λ_j and t^k cannot be separated for a given pair, the ratio for two different values of j should be roughly constant across k. The AB method utilizes this property through least-squares to estimate the relative eigenvalues of Λ. One of the eigen values is set to 1 and then, up to a scalar multiple, the rate matrix R is taken as $\hat{U}\hat{\Lambda}\hat{U}^{-1}$, where $\hat{U}$ is the estimate of U; $\hat{\Lambda}$ is the estimate of

Λ. This scalar multiple for R is chosen so that $-\sum_j \hat{\pi}_j \hat{R}_{jj} = 1$, where $\hat{\pi}_j$ is the estimate of stationary frequency of j^{th} character base; $\hat{R}$ is the estimate of R. This is the conventional rescaling to ensure that edge lengths are interpretable as expected numbers of substitutions.

1.3 Nonstationary Models

The assumption of stationary of evolutionary processes along edges is very restrictive since it requires that entries in character state frequency vectors at all nodes remain the same up to random error. Stationary models may or may not, therefore, adequately describe true evolutionary processes. Indeed the stationarity assumption is well known to be violated in cases of compositional heterogeneity across different sequences (Bernardi 1993, Foster 2004, Foster and Hickey 1999, Foster et al. 1997, Hasegawa and Hashimoto 1993, Jukes 1986, Lockhart et al. (1992, 1994), Montero 1990, Mooers and Holmes 2000, Yang and Roberts 1995). Unfortunately, the use of stationary models in estimation when nonstationary processes are operating may lead to biased phylogenetic estimates such as the artefactual grouping of species with similar nucleotide frequencies together even if these organisms are not close relatives. To avoid this problem, models have been developed which take into account the differences in base frequencies among species. Nonstationary models such as Galtier and Gouy (1995, 1998), referred to as the GG model, Yang and Roberts (1995), referred to as the YR model, Barry and Hartigan (1987), referred to as the BH model, have been developed to overcome compositional bias and produce more accurate results (e.g., see Sheffield et al. (2009)).

The GG and YR models are two commonly used non-stationary models. In contrast to time-reversible models, and as with most non-stationary models, probabilities of data differ depending upon where the root of the tree is. Thus analyses with these models must specify a root or attempt to estimate its location. In these two models, the root frequencies normally are not the same as the stationary frequencies along the edge connecting the root and an internal node and each edge has its own stationary model and edge length. The GG and YR models were developed with one goal being the reduction of model complexity and thus had simple stationary models along edges. In the YR model, the stationary model of each edge was derived from the HKY85 model (Hasegawa et al. 1985) in which the stationary frequency vector had no constraints other than to sum to one, and the edges share the same transition and transversion ratio. As with stationary models, edge lengths were defined as expected numbers of substitutions under the stationary model for the edge. In the GG model, which is a special case of the YR model, the stationary model of each edge was derived from the T92 (Tamura 1992) which constrains the frequencies of character states C and G to be the same and constrains the frequencies of characters A and T to be the same.

The BH model (Barry and Hartigan 1987) assumes that: 1. all sites are independent; and, 2. from an internal node, the evolutionary processes from the parent to daughters are independent Markov processes. For instance, the nodes 2 and 1 in Figure 1.1a are both the children of node 3. The evolutionary process along the edge (3, 1) is independent on the process along the edge (3, 2). The BH model differs from most phylogenetic models in that the substitution matrix along an edge need not correspond to a continuous-time Markov process model. It only needs to satisfy the constraints of a discrete time transition matrix. This allows the BH model to model more complicated processes that do not correspond to any continuous-time model. An example is when there is a zero probability of substitution of one (or more) particular character states to another. Furthermore, in the BH model, each edge can have a completely different process associated with it. For the tree shown in Figure 1.1a, the evolutionary processes are different along edge $(a, 3)$, $(3, 1)$, $(3, 2)$ and etc. Also the evolutionary processes of two directions of an edge, e.g., the directions from node a to node 3 and from node 3 to node a, are not the same in general. For the YR and GG models, the parameters for any edge are the stationary frequency vectors. In contrast, the parameters for any edge of a BH model are the joint probabilities of different character states at the two nodes. Using internal consistency, we can show that the likelihoods of a BH model will remain the same no matter which data node is selected as the root, a convenient property when doing estimation. Thus the estimates of the BH model can be estimated using an unrooted tree which is a clear advantage by contrast with the need to consider rooted trees for estimation in other nonstationary models such as the GG and YR models. As shown in Jayaswal et al. (2005), estimates of frequencies at terminal nodes are the same as the empirical terminal frequencies, which is a nice property. However, since processes along edges of a BH model are not required to be continuous-time Markov processes, the model does not give edge lengths for a tree. This model was reconsidered and implemented most recently by Jayaswal et al. (2005). As a further development of the BH model, Jayaswal et al. (2011) proposed two stationary and non-homogeneous models as simplified versions of the BH model.

1.4 Rate-across-sites Variation

1.4.1 Stationary Models

It has long been recognized the evolutionary dynamics of different sites in a molecule differ depending a variety of factors. For instance, in Fitch and Margoliash (1967), invariant sites were identified in the amino acid sequences of *Cytochrome C*. These sites were thought to be fixed at the particular amino acid by purifying selection (i.e., any non-synonymous changes at these sites would cause the resulting protein to cease to function properly and the resulting mutant would have a strongly negative fitness effect). Such rates-across-sites variation also happens in the different codon positions where the third codon position often has the fastest rate and the second codon position has the slowest rate (Yang 1996). As pointed out by Semple and Taylor (2009), the physical structure of genome can also affect the rate of substitutions.

Assume that the evolutionary rate for the k^{th} site is r_k. In a continuous-time Markov model, the probability of transition from state i to state j separated by the evolutionary time t^{ij} becomes $P(r_k t^{ij})$ instead of $P(t^{ij})$. As previously mentioned, the evolutionary processes among sites are usually treated as independent and share a rate distribution. Site rate distributions previously used in phylogenetic analyses include the gamma distribution (Jin and Nei 1990, Nei et al. 1976, Uzzell and Orbin 1971, Yang 1993), a log normal distribution (Olsen 1987), an inverse normal distribution (Waddell and Steel 1997), or discrete rate classes in proportions (Hasegawa et al. 1987). Taking the invariable site as a category and separating it from others, Gu et al. (1995) proposed a rate distribution mixture model of a discretized gamma plus invariable site, $\Gamma + I$, which has subsequently been incorporated into applications such as PAML(Yang 2007), PAUP* (Wilgenbusch 2003), PHYLIP (Felsenstein 1989) and RAxML (Stamatakis 2006), amongst others.

For commonly used Γ distribution of site rates as indicated in Yang (1993), the scale factor β is set to be equal to the shape parameter α so that the resulting distribution has mean 1; a mean of 1 ensures that edge-lengths are interpretable as expected numbers of substitutions. For a site $\underline{x}$, the distribution with considering the site rates distribution is $P(\underline{x}|\Theta) = \int_{r=0}^{\infty} P(\underline{x}|\Theta, r)p(r)dr$, where $p(r)$ is the gamma rate probability density function; Θ is the set of parameters (cf. Chapter 4 in Yang 2006). Because of the integration, the pruning algorithm cannot be applied to $P(\underline{x}|\Theta)$ making exact computation infeasible with a large number of taxa. Yang (1994b) has provided an approximate solution by assuming a discretized rate distribution based on gamma distribution for the rate variable. The distribution of the rates is separated into several equiprobable rate intervals. Then the mean or median of each interval is taken as the estimates of the rates. In practice 4-8 intervals are frequently used.

In model-based analyses of sequences, it is not a common practice to check whether the evolutionary rates among site are variable prior to applying a model. However, modeling of different rates at different sites is important because studies in Fitch and Margoliash (1967), Jin and Nei (1990), Kuhner and Felsenstein (1994), Lockhart et al. (1996), Shoemaker and Fitch (1989), Steel et al. (2000), Song (2010), Wakeley (1994), Yang (1994b, 1995, 1996) have shown that the estimates of topologies and model parameters will not be recovered properly if the rates among sites are not the same but these differences are ignored.

1.4.2 Incorporating Site-rate Variation into Nonstationary Models

For both nonstationary models, the YR and GG models, with continuous Markov processes along edges, it is easy to incorporate a gamma distribution to allow for rate variation. Gamma models have been considered in both of the YR model implemented by PAML (Yang 2007) and the GG model implemented by nhPhyML (Boussau and Gouy 2006). However neither of them has considered invariable sites as part of their models.

For the BH model, incorporating site-rate variation is less straightforward. Jayaswal et al. (2007) proposed the BH + I model that explicitly accounts for invariable sites on the top of the BH model. This is a special two-class solution to rates-among-sites variation by setting one class having a zero rate and the other class as sharing the same rate among sites. The simplified version of the BH models in Jayaswal et al. (2011) also allow invariable sites. However, incorporating a gamma distribution correction for site rates into the BH model is not straightforward because it lacks edge lengths which in standard phylogenetic models serve as the parameters that are rescaled by rate multipliers coming from a discretized gamma distribution. In Chapter 5, we address this problem by a novel mixture model approach to modeling site rates and other site-specific variations in the evolutionary process in the context of the BH model.

1.5 Identifiability in Phylogenetic Analysis

A nonidentifiable model in statistical analysis is one where there exists more than one set of parameters that give the same distribution (Bickel and Doksum 2007). If a phylogenetic model is nonidentifiable, there exists more than one set of parameters for that model which give the same site pattern distribution. Let $P(\underline{x};\Theta)$ denote the distribution of the site pattern $\underline{x}$. Here, $P(\underline{x};\Theta)$ is the joint probability of $\underline{x}$ under the model with the parameters Θ. If Θ is nonidentifiable, it is clear that one cannot distinguish between different Θs which gives the same site pattern distribution, no matter how much data is collected.

Later in Chapter 3, we will consider a more complicated example of nonidentifiability in phylogeny. For illustration, as a simpler example, consider estimation of the time since divergence for two taxa. The joint probability with which the nucleotides i and j are observed is $\pi_j P_{ij}(2rt)$ where r is the rate at which substitutions occur and t is the time since divergence. Assume the rate is 10^{-9} (e.g., in Hasegawa et al. 1985). If the time since divergence for the two taxa was 200 Myr, the joint probability distribution would be $\pi_i P_{ij}(0.2)$. However this joint distribution is the same as the one that arises when $r = 5 \times 10^{-9}$ and $t = 40$Myr. Since only the joint probability of i and j can be inferred from data, one can never distinguish between these two values (r, t). What can be estimated, however, is the product $rt = 0.2$. It is the expected number of substitutions that took place since divergence of the two taxa.

Assuming edge lengths are interpreted as expected numbers of substitutions, thus avoiding the identifiability issue alluded to above, identifiability results have been established in Chang (1996). There, in Theorem 4.1, it has shown that if estimates are constrained to be within a class of transition matrices that is reconstructible from rows, those matrices are identifiable. This theorem was proved without any constraints on the transition matrices along edges. This does not, however, imply that all parameters of models that give rise to transition matrices are identifiable. Moreover, the result in Chang (1996) does not deal with rates-across-sites models. Allman et al. (2008) proved that for some models, not all parameters in general are identifiable. Chang (1996) also gives some examples of transition matrices that are not reconstructible from rows.

Some authors have considered a looser definition of identifiability, referred to as generic identifiability. A model is considered generically identifiable if the set of parameters that is not identifiable is of zero Lebesgue measure. Theorem 1 of Allman et al. (2008) has proved that the GTR+Γ is identifiable for all parameters when there are four character states. However for other number of states, the

GTR+Γ model is only "identifiable from the joint distributions of triples of taxa for generic parameters on any tree with three or more taxa." The implication is that there is a relatively large subspace of parameters that can be identified with enough data. It remains possible, however, that there will be some parameter which, if they happen to be the generating parameters, are nonidentifiable. Usually generic identifiability results do not characterize the set of nonidentifiable parameters and thus allow for the possibility, albeit unproven, that all parameters are identifiable.

Rogers (2001) argued that the GTR + Γ + I model is identifiable. Even though Allman et al. (2008) suggest that the GTR + Γ + I model is identifiable for generic parameters, they pointed out gaps in the proof of Rogers (2001). Recently, Chai and Housworth (2011), proved that for generic parameters, the GTR+Γ +I model is identifiable as claimed in Rogers (2001).

The BH estimates are joint probability matrices along edges. As we mentioned before, the likelihood for a fix topology doesn't depend upon the rooting position. Taking node a in Figure 1.1a as the root, we can write the likelihood of site pattern $\underline{x} = \{x_a, x_b, x_c, x_d, x_e\}$ as the following.

$$P(\underline{x}) = \sum_{x_1} \sum_{x_2} \sum_{x_3} \frac{G(X)}{\pi_{x_1}^2 \pi_{x_2}^2 \pi_{x_3}^2} \tag{1.3}$$

where $G(X) = F_{a,3}(x_a, x_3) F_{3,1}(x_3, x_1) F_{3,2}(x_3, x_2) F_{1,b}(x_1, x_b) F_{1,c}(x_1, x_c) F_{2,d}(x_2, x_d) F_{2,e}(x_2, x_e)$; $F_{ij}(x_i, x_j)$ are the joint probability which observes a state x_i at node i and a state j at node j; π_{x_j} is the frequency of the state x_j at node j. Relabelling the nucleotides at nternal nodes 1, 2 and 3 has no effect on the site pattern distribution in (1.3) which can be extended to site pattern distributions for any number of taxa. Since (1.3) shows that more than one set of the parameters - which in this case are joint probability matrices along edges - can give the same site pattern distributions, the estimates of the BH model are not identifiable. This problem was been identified in Barry and Hartigan (1987) and independently by us.

Although it does not directly utilize parameters of the BH model, the logDet distance introduced by Steel (1994) and Lockhart et al. (1994) can be used with distance methods like the Neighbor-Joining method of Saitou and Nei (1987) to recover the topology. The logDet distance is the logarithm of the determinant of a frequency matrix for two species, a and b; the ij^{th} entry of the frequency matrix is the number of occurrences of the i^{th} character state for species a and j^{th} character state for species b. As shown in Steel (1994), if the joint probability matrices of data are used as frequency matrices, the resulting logDet distances are tree-additive. It follows that logDet distances can be used to consistently recover the tree topology when a BH model generates the data.

1.6 Data Sets

In this thesis, the PANDIT database of Whelan et al. (2006) and the *Plasmodium* data set of Davalos and Perkins (2008) were used. The PANDIT database of Whelan et al. (2006) was used in Chapter 2 to test our new methods for estimation of rate matrices for amino acid analysis, which is comparable with the PAM, JTT and WAG models. The *Plasmodium* data set was used for our work on the non-stationary models in Chapters 3, 4, and 5.

The PANDIT database of Whelan et al. (2006) had 7738 protein families and 174,760 sequences. Those widely ranging sequences were based on the collection of the Pfam database of Bateman et al. (2004). The alignments are available for both amino acid and DNA sequences for each protein family. The database also contains the estimated tree for each protein family.

The data set of Davalos and Perkins (2008) contains eight DNA sequences from eight species of the genus *Plasmodium: P. berghei, P. chabaudi, P. falciparum, P. gallinaceum, P. knowlesi, P. reichenowi, P. vivax, P. yoelii.* The data set was constructed by using the genes in the *P. falciparum* for which homologous sequences could be identified in seven other *Plasmodium* species but excluding all within-genome duplicate copies as identified in the OrthoMCL database (Chen et al. 2006). The data set contains 104 genes selected from *Plasmodium* genome. The alignment with gaps removed has 77313 sites. Davalos and Perkins (2008) found that G+C contents for *P. knowlesi* and *P. vivax* were significantly higher than for the rest. These data were chosen for our investigations of non-stationary models.

1.7 Overview of Future Chapters

In Chapter 2, we present a new method to construct an instantaneous rate matrix for amino acid data. It deals with the difficulties of sparseness for amino acid data. As mentioned before, the methods by which the PAM, JTT and WAG models were derived all have shortcomings. For example, if we want to construct a rate matrix from a large database, we have to find a way to overcome the constraint of 85% or higher similarities required by the methods used to build the PAM and JTT models and the high computational cost of ML employed for the estimation of the WAG matrix.

The AB method estimates an instantaneous rate matrix from pairs of sequences of DNA data sets without doing optimization and constructing a tree. However, we found that this method does not work well for an amino acid data set because of the sparse pairwise transition matrices obtained (Zou 2005). In this thesis, we propose a method of estimating a rate matrix for amino acid datasets that extends the approach of Arvestad and Bruno (1997) but which solves the problem of lack of data. We will introduce a binning method which will replace the pairwise comparison of the AB method. Using our method, it is possible to estimate a rate matrix for a single protein family or for a database that contains multiple protein families. The rate matrix we estimated from the PANDIT database of Whelan et al. (2006) is comparable with existing models such as the PAM , JTT, and WAG.

The rest of the thesis considers extensions of the BH model. Initial investigation in Chapter 3 leads to the conclusion that the parameters of the BH model are not all identifiable. This happens because permutations of the frequencies at internal nodes give the same site likelihood. Thus, internal node frequency vectors and edge lengths cannot be directly recovered.

In Chapter 4, we present an algorithm for accurately estimating the correct permutations of the BH estimate and a formula for calculating an approximate edge lengths for the BH model. In order to estimate the correct permutations, we define nonstationary GTR model, referred to as the NSGTR model. The NSGTR model has a GTR model along each edge but allows these models to vary along edges. To estimate correct permutations for the BH model, for each edge and each direction, a GTR model is determined that best fits transition matrix along the edge and the direction. Since all GTR parameters are identifiable (Allman et al. 2008), this method can consistently estimate internal node frequencies and edge lengths if a NSGTR is applicable.

Minin and Suchard (2008a) have given formulas for expected numbers of substitutions for a nonstationary continuous-time Markov process. Since our NSGTR method gives estimates of the continuous-time Markov models for all estimates of a BH model, we utilize these formulas to estimate edge lengths for a BH model fit. In the implementations of the YR and GG models, edge lengths are estimated in the conventional manner for stationary and reversible models which is not appropriate for a nonstationary model. Our method can also be used with the YR and GG methods to obtain more appropriate nonstationary edge length estimates.

We used simulations to test our methods estimating permutations and edge lengths. In our simulations, we used the estimates of the NSGTR models for the *Plasmodium* data set, which we called the “mild” parameters in the sense that there does not exist extremely large or small GTR parameters, and a set of parameters that contain extremely large and small GTR parameters. The simulation results showed that regardless of whether mild or extreme parameters were employed, our algorithm estimated the correct permutations. The recovered edge lengths were also very close to the generating edge lengths. We are confident that our method is useful in some cases for finding interpretable edge lengths and internal node frequency vectors in nonstationary models. We conclude the chapter with an analyses of the *Plasmodium* data set of Davalos and Perkins (2008).

In order to allow for differing processes at different sites as in rates-across-sites variation, we propose a BH mixture model in Chapter 5. Our BH mixture model does not assume a specific distribution for the evolutionary rates of classes among sites. In our solution, each class in the BH mixture model is either a regular BH model or a reduced model for invariable sites. With two classes in which one of these two classes is invariable sites, it becomes the BH + I model of Jayaswal et al. (2007); with K classes and one of classes being invariable sites, we refer to it as the BH($K-1$) + I model; with K variable classes, we refer to it as the BHK model. Identifiability issues still exist for the estimates of any given class except for the estimates for the invariable sites. Therefore if the internal frequencies and edge lengths are of interest, the permutations have to be estimated via the NSGTR methodology developed in Chapter 4. Simulations showed solid estimation performance for NSGTR and BH mixture models. Our simulation results also showed that the BH mixture model is useful for modeling compositional heterogeneity and better fits real data than the simple BH or BH+I model.

Chapter 2

Constructing a Rate Matrix for Amino Acid Evolution from Pairs of Sequences

2.1 Introduction

Instantaneous rate matrices play an important role in reconstructing phylogeny under the general time-reversible (GTR) Markov model. Estimating a rate matrix directly for an amino acid data set of interest may in some cases be more appropriate than using a fixed rate matrix such as the JTT. Unfortunately, estimating an instantaneous rate matrix using an amino acid data set is more difficult than using a DNA data set because 209 parameters must be estimated under the time reversible Markov model in contrast to 9 parameters with a DNA data set. Moreover the length of an amino acid sequence is shorter than the length of the corresponding DNA sequence.

Three of the fixed rate matrices commonly used with amino acid data are the PAM (Dayhoff et al. 1979), JTT (Jones et al., 1992) and WAG (Whelan and Goldman, 2001). Among these, both the PAM and JTT matrices are estimated using the parsimony method. The shortcoming of the parsimony method is that it assumes that an observed difference in character states is the result of one substitution. As a result, it uses the pairs in a data set which have 85% or higher simularities; the consequence is the loss of a large amount of information from the original data sets. The advantage of this method is that it is simple and computationally efficient. The WAG method uses maximum likelihood (ML). In this approach, to avoid the computational cost of full ML estimation, the topology of each protein family in the data set was constructed using the neighbor-joining method (Saitou and Nei, 1987) based on the pairwise distances obtained under the Dayhoff+F[1] model. After constructing the topologies for all protein families, all edge lengths were re-estimated under the JTT+F model by the ML method. Then, the rate matrix was obtained through ML with fixed topologies obtained from the neighbor-joining method and fixed ratios of edge lengths across alignments. The WAG rate matrix is attractive in the sense that the ML estimation is statistically sound whereas parsimony has undesirable statistical properties (cf. Chapter 9 of Felsenstein 2004). However, if the topologies are far from correct, the estimated rate matrix may be inaccurate.

In contrast to these previous methods, Arvestad and Bruno (1997) proposed a method, which we refer to as the AB method, for estimating the instantaneous rate matrix of a DNA data set using pairwise distance information without having any

[1] The suffix $+F$ means that the estimates of the stationary frequencies are the proportions of the amino acids as observed in the data set of Whelan and Goldman (2001)

constraint on the similarity of sequences and without requiring the estimation of tree topologies under the GTR model. This method is attractive because it eliminates the similarity constraint of counting methods and avoids the heavy computational cost of ML.

The AB method uses empirical estimates of the transition matrices from pairs of taxa. In theory, the eigenvectors of the transition matrices are the same as the eigenvectors of the rate matrices. The AB method estimates these through the eigenvectors of the average estimated transition matrices. In theory, the logarithms of the eigenvalues of the transition matrices are proportional to the eigenvalues of the rate matrix; for a given pair, the constant of proportionality is the evolutionary distance for the pair. The AB method uses logarithms of eigenvalues from each estimated pairwise transition matrix and a least-squares approach to estimate the eigenvalues of the rate matrix.

Pairwise amino acid data does not always provide enough information to allow the AB method to accurately estimate the 209 parameters of the rate matrix and the base character frequencies. We have noticed several such problems. For instance, when estimating a rate matrix using the software implementation of Arvestad and Bruno (1997), we found that the estimate of the rate matrix changes if the order of the sequences in a data set is changed. Also, it is not uncommon that some estimated eigenvalues are negative, that the estimated instantaneous rate matrix has negative off-diagonal elements or that the sums over the rows of the estimate of the rate matrix are not equal to zero. In other words, the estimates of the rate matrices do not satisfy the properties of a rate matrix.

To overcome this lack of information, we construct bins over the proportions of the pairwise differences in a data set and assign the pairs into bins. Two pairs of sequences are assigned to a bin if they have a similar proportion of pairwise differences. In comparison to using a pair as a unit to extract the distance information in the AB method, we use a bin as a unit.

2.2 The AB Method

2.2.1 Overview

Define $P(t^k)$ as the transition matrix of the k^{th} pair which has the evolutionary distance t^k. Suppose $P(t^k)$ has eigenvectors decomposition $Ue^{\Lambda t^k}U^{-1}$, where the eigenvalue matrix Λ is a diagonal matrix and U and U^{-1} are the right and left eigenvectors; note that since the rate matrix is assumed the same for each pair, eigenvectors of $P(t^k)$ are the same for each pair. As shown in the equation 17 of Arvestad and Bruno (1997), which is reproduced below, $\lambda_j t^k$ can be calculated by

$$\lambda_j t^k = \log u_j^\dagger P(t^k) u_j \tag{2.1}$$

where $j \in \{1, \ldots, 4 \text{ or } 20\}$ for DNA or amino acid sequences; t^k is the evolutionary time between the two taxa in the k^{th} pair; λ_j is the j^{th} eigenvalue of the rate matrix; $u_j^\dagger$ and u_j are the j^{th} row and column of the left eigenvectors, U^{-1}, and the right eigenvectors, U, of the rate matrix. In the AB method, estimates $\hat{P}^k$ of $P(t^k)$ are obtained from observed pairwise frequencies for the k^{th} pair and estimates of the j^{th} elements of U and U^{-1}, $\hat{u}_j$ and $\hat{u}_j^\dagger$, are obtained from the eigenvector decomposition of $\frac{1}{n}\sum_k \hat{P}^k$. Substituting $\hat{u}_j$, $\hat{u}_j^\dagger$ and $\hat{P}^k$ in (2.1) gives an estimate $\hat{\lambda}_j^k$ of $\lambda_j t^k$.

2.2.2 Improvements to the AB Method

2.2.2.1 Estimation of the Frequency Matrix

One improvement of the AB method comes from using a different estimate of $P(t^k)$. Let $\hat{F}^k$ be the estimate of the frequency matrix for the k^{th} pair; its ij^{th} entry is the frequency with which the state i is observed for the first sequence and the state j for the second. In the AB method, the matrix $\hat{P}^k$ is obtained by rescaling the entries of $\hat{F}^k$ so that rows sum to 1. We have observed that the resulting estimates of the rate matrices, differ if the orders of the pairs of sequences are changed. This is because, due to random variation, $\hat{F}^k$ is rarely symmetric even though, under the GTR model, it is estimating entries of a symmetric joint probability matrix. Replacing it with $\hat{F}_s^k = [\hat{F}^k + (\hat{F}^k)^T]/2$ gives a valid estimate of the joint probability matrix and avoids having results that depend upon the order in which sequences were listed in an alignment.

2.2.2.2 A Symmetrized Version of $P(t^k)$

Arvestad and Bruno (1997) show that $P(t^k)$ can be replaced by a symmetrized version, $P_s(t^k)$, in the key equation (2.1) leading to the AB method. The reason that replacement is worthwhile in practice is that it ensures that eigen-decomposition routines specifically designed for symmetric matrices can be used. We utilize the same approach here but with some modification. The symmetrized $P(t^k)$ are

$$P_s(t^k) = (\Pi^k)^{-\frac{1}{2}} * F(t^k) * (\Pi^k)^{-\frac{1}{2}} = (\Pi^{-\frac{1}{2}} * F(t^k) * \Pi^{-\frac{1}{2}})^T = \Pi^{\frac{1}{2}} * P(t^k) * \Pi^{-\frac{1}{2}} = U_s e^{\Lambda t^k} U_s^T$$

where the diagonal matrix Π^k gives the stationary frequencies of $P(t^k)$; the diagonal matrix Π is the average Π^k. These are assumed equal across pairs so that $\Pi^k = \Pi$. It isn't difficult to show that the eigenvectors of $P_s(t^k)$, U_s, satisfy $U_s = \Pi^{\frac{1}{2}} U$ and $U_s^{-1} = U^{-1}\Pi^{-\frac{1}{2}}$ and the eigenvalues of $P_s(t^k)$ are the same as the eigenvalues of $P(t^k)$. It follows that $\tilde{P}_s = (1/n)\sum_k P_s(t^k) = (1/n)U_s \sum_k e^{\Lambda t^k} U_s^T$ is symmetric and has the same eigenvectors as $P_s(t^k)$. In practice, we estimate $\tilde{P}_s$ by $\bar{P} = (1/n)\sum_k \hat{P}_s^k$ where $\hat{P}_s^k = (\hat{\Pi}^k)^{-\frac{1}{2}} \hat{F}_s^k (\hat{\Pi}^k)^{-\frac{1}{2}}$ and, for the k^{th} pair, $\hat{\Pi}^k$ is a diagonal matrix with the observed frequencies of the observed frequencies of the amino acids along its diagonal. From the eigen-decomposition $\bar{P}$, we obtain the estimates of U_s and U_s^T, $\hat{U}_s$ and $\hat{U}_s^T$, where $\hat{U}_s$ has the j^{th} column $\hat{u}_{sj}$ and $\hat{U}_s^T$ has the j^{th} row $\hat{u}_{sj}^\dagger$. Using $\hat{U}_s$ and $\hat{U}_s^T$ gives the estimates $\hat{U}$ and $\hat{U}^{-1}$ of right and left eigenvectors matrices U and U^{-1} through $\hat{U} = \hat{\Pi}^{-\frac{1}{2}} \hat{U}_s$ and $\hat{U}^{-1} = \hat{U}_s^T \hat{\Pi}^{\frac{1}{2}}$, where $\hat{\Pi}$ is the estimate of Π obtained by averaging $\hat{\Pi}^k$. For the GTR model, $u_{sj}^\dagger P_s(t^k) u_{sj}$ is the same as $u_j^\dagger P_s(t^k) u_j$, so that symmetrized or unsymmetrized values can be used interchangeably in (2.1). Thus valid estimates of $\lambda_j t^k$ can be obtained by substituting $\hat{u}_{sj}^\dagger$, $\hat{u}_{sj}$ and $\hat{P}_s^k$ in (2.1). This is what is suggested in the AB method. A difficulty with this approach is that it is not guaranteed that the largest eigenvalue will be 0 that is required by a rate matrix. Thus as an estimate of $\lambda_j t^k$ we use the following equation.

$$\hat{\lambda}_j^k = \log \frac{\hat{u}_{sj}^{\dagger} \hat{P}_s^k \hat{u}_{sj}}{\hat{u}_{s1}^{\dagger} \hat{P}_s^k \hat{u}_{s1}} \tag{2.2}$$

For highly divergent pairs, $\hat{u}_{sj}^{\dagger} \hat{P}_s^k \hat{u}_{sj}$ may be negative which prohibits the use of (2.2). In practice, with eigenvalues in descending order, we replace each negative estimate by the first nonnegative estimate $\hat{u}_{sk}^{\dagger} \hat{P}_s^k \hat{u}_{sk} > 0$ which satisfies $k > j$. In our experiments, we have found that replacing $\hat{u}_j^{\dagger}$, $\hat{u}_j$ and $\hat{P}^k$ by $\hat{u}_{sj}$, $\hat{u}_{sj}^{\dagger}$ and $\hat{P}_s^k$ tends to reduce the number of negative off-diagonal elements in an estimate of the rate matrix.

2.2.2.3 Rescaling the Rate Matrix

Since $\hat{\lambda}_j^k$ in (2.2) estimates $\lambda_j t^k$, $\hat{U}\hat{\Lambda}^k\hat{U}^{-1}$ estimates $R^k = U\Lambda t^k U^{-1}$. Let $R^* = \sum_{k=1}^K R^k = U\Lambda U^{-1}(\sum_{k=1}^K t^k)$, where K is the total number of pairs. Then an estimate of R^* is obtained as

$$\hat{R}^* = \sum_{k=1}^{K} \hat{U}\hat{\Lambda}^k(\hat{U})^{-1} \tag{2.3}$$

Normally the rate matrix is rescaled so that the average substitution rate per site is equal to one. Since this rescaling replaces R by $-\frac{R}{\sum_j \pi_j R_{jj}}$ and since $\frac{R^*}{\sum_j \pi_j R_{jj}^*} = \frac{R}{\sum_j \pi_j R_{jj}}$, an estimate of the rate matrix from K pairs is

$$\hat{R} = -\frac{\hat{R}^*}{\sum_j \hat{\pi}_j \hat{R}_{jj}^*} \tag{2.4}$$

Our approach to adjusting for the presence of t^ks is simpler than the AB method which uses a complicated least squares method.

One of the difficulties we encountered with the AB method software was that row sums of R were not always 0. This is not a problem with our method. For instance, using our modifications above, we estimated a rate matrix for the data set in Yang (1994a). Our results showed that our estimate of the rate matrix has row sums in the range $(-9.66E-15, 9.44E-15)$. We compared entry by entry of the estimates of using our method with the ML estimates in Yang (1994a) and calculated the relative changes between our estimates and the estimates in Yang (1994a). For the sixteen entries, the minimum, first quantile, median, third quantile and maximum, called the five-number summary, of those relative changes are $(0.0004, 0.002, 0.017, 0.021, 0.095)$. A similar comparison for the AB method gives the five-number summary (0.014, $0.033, 0.048, 0.068, 0.15$). Since ML uses more information and is expected to give more precise R estimates, a clear improvement has been shown with our modifications of the AB method. Similar improvements occurred in the resulting estimates of $P(t)$ for varying values of t.

2.3 Binning Methods

2.3.1 Difficulties Using the AB Method for Amino Acid Data Sets

To illustrate the difficulties with the AB method applied to amino acid data sets, we consider two data sets in the PANDIT database of Whelan et al. (2006). With gaps removed, they have sequence lengths of 2096 and 1241 sites with 4 and 18 taxa respectively. We found that there were up to 132 zero entries in the empirical pairwise transition matrices constructed from these two data sets; $P(t^k)$ should have none. In addition, some of the eigenvalues of the $\hat{P}^k$s were negative; $P(t^k)$ has all positive eigenvalues. As a result, we were not surprised to see negative off-diagonal entries in the estimate of the rate matrix. This observation strongly suggests that pairs of sequences with 2000 sites do not have enough information for accurate estimation with the AB method. The sequence lengths of the majority of protein families in the PANDIT database of Whelan et al. (2006) are much less than 2000 sites. If we want to utilize the advantages of the AB method, we have to determine a mechanism to deal with this lack of information.

2.3.2 Binning Sequences

Using similar ideas as in Henikoff and Henikoff (1992), we propose a new unit, a bin, that groups pairs of sequences that have the same or similar pairwise evolutionary distances, thus increasing the information in each observational unit. When pairs are close enough, the differing distances of the pairs in a bin can be ignored and a “super” pair can be constructed by concatenating all pairs. If there is only one pair in each bin, this is equivalent to the AB method. By putting enough pairs in a bin, the difficulties of sparse transition matrices alluded to above can be avoided.

Let $\mathcal{K}_b$ label the pairs in the b^{th} bin; let m_b be the number of pairs in the b^{th} bin. For the j^{th} pair in the b^{th} bin, $F(t^j)$ is the joint probability matrix. Taking the average over all pairs in the bin, we define

$$P^b = \frac{1}{m_b}\sum_{j\in\mathcal{K}_b} P(t^j) = \frac{1}{m_b}\sum_{j\in\mathcal{K}_b} \Pi^{-1}F(t^j) = \frac{1}{m_b}U\sum_{j\in\mathcal{K}_b} e^{\Lambda t^j}U^{-1} \tag{2.5}$$

Since the t^j are approximately the same for all members of b, P^b is approximately the transition matrix for any one of the pairs. Let λ_i^b be the i^{th} element of the logarithm of the eigenvalues of P^b. In our binning approach, $\tilde{P}^B = \sum_b P^b$, where B is the number of bins, plays the role that $\tilde{P}$ played without binning. Estimation of R is as described in previous subsections without binning but replacing $\hat{P}^k$, $\hat{\Pi}^k$ and $\bar{P}$ with the estimates $\hat{P}^b$, $\hat{\Pi}^b$ and $\bar{P}^B$ of P^b, Π^b and $\tilde{P}^B$. Similarly as with the pairwise method in previous section, the symmetrized version, $P_s^b = \Pi^{\frac{1}{2}}P^b\Pi^{-\frac{1}{2}}$, has the same eigenvalues as P^b. In the pairwise method, replacing $\hat{u}_j^\dagger$, $\hat{u}_j$ and $\hat{P}^k$ by $\hat{u}_{sj}$, $\hat{u}_{sj}^\dagger$ and $\hat{P}_s^k$ improves the estimate of the rate matrix by tending to reduce the number of negative off-diagonal elements. This is also true for binning upon replacing $\hat{P}^b$, $\hat{u}_j^\dagger$ and $\hat{u}_j$ by $\hat{P}_s^b$,$\hat{u}_{sj}^\dagger$ and $\hat{u}_{sj}$; the symmetric matrix $\hat{P}_s^b$ is calculated as before but replacing pairwise quantities by their analogues for bins. Replacing $\hat{P}^k$ by $\hat{P}^b$ in (2.2) gives the estimate $\hat{\lambda}_j^b$ of j^{th} eigenvalue of P_s^b. Using $\hat{U}$, $\hat{U}^{-1}$ and $\hat{\Lambda}^b$, where the diagonal matrix $\hat{\Lambda}^b$ has the diagonal elements $\hat{\lambda}^b$, gives $\hat{R}^* = \sum_b \hat{U}\hat{\Lambda}^b\hat{U}^{-1}$. The rescaling formula in (2.4) can still be used to obtain an estimate of the rate matrix, $\hat{R}$.

2.3.3 Bin Construction

Bin construction is an important part of the procedure. If bins are too large, P^b will provide a poor approximation to the transition matrices for the bin. If bins are too small, P^b will be sparse. We propose two binning strategies: equal size and equal width. For either strategy, a pair of sequences is assigned to a bin if the interval of this bin includes the estimated evolutionary distance for the pair. Using the equal size binning strategy, intervals are chosen so that all bins have the same number of pairs. Under the equal width strategy, each bin has the same interval length. The remaining choice for bin construction is the number of bins. We investigate what usually are good choices through simulations.

2.4 Simulations

We carried out simulations to assess the possible binning strategies. In our simulations, we used the JTT model of Jones et al. (1992) as our true model. All data sets were simulated using seq-gen (Rambaut and Grassly 1997). When exploring the simulation results, we focused on analyzing the average mean squared errors and the average biases of the estimates of eigenvalues and rate matrices. We set up two types of simulations: one based on the evolutionary distances, simulation 1; the other on the proportion of pairwise differences, simulations 2, 3, and 4. Since our method only requires pairwise distances, we simulated the pairs over various distances without considering the topology.

In simulation 1, we studied the effect of evolutionary distances and sequence lengths. Simulations under a JTT model were separately conducted for each of the 10 distances, $0, 0.5, 1.0, \ldots 4.5$. In each simulation, 500 pairs of sequences were independently generated from the given distance to obtain estimates of rate matrix parameters. To investigate the effect of sequence lengths, simulations were repeated with 200, 500 and 2000 sites. For each simulation setting (distance and number of sites) results were obtained for 100 simulations.

Simulation 2 was designed to test the performance of our algorithm on a large data set. To ensure that sampled distances for pairs approximated those that one might encounter in practice we utilized the PANDIT database of Whelan et al. (2006) as follows. A large set of sequences was sampled and the proportions of differences for each pair of sequences was obtained. Each proportion, p, was converted to the expected number of substitutions t by solving

$$p = 1 - \sum_j \pi_j P_{jj}(t) \tag{2.6}$$

where the π_j is given in the JTT model and $P_{jj}(t)$ were calculated under the JTT model. Doing this for pairs of sequences coming from the PANDIT database gave a distribution of distances. Evolutionary distances were generated from this distribution.

We had other two simulations, simulations 3 and 4, for evaluating our procedure for small data sets. All evolutionary distances in these two simulations were sampled and generated from the PANDIT distance distribution described above. In simulation 3, the number of taxa in a data set was fixed while the number of sites was a variable; in simulation 4, the sequence length in a data set was fixed while the number of taxa was a variable.

2.5 Results and Discussion

Sequence length is an important factor in our binning strategy. As a first step, we determine a minimum sequence length which will be used to investigate binning strategies in other simulations. Figure 2.1, based on simulation 1, shows the mean squared errors of the estimates of the second largest eigenvalue λ_2 and the smallest eigenvalue λ_{20}. The estimate of the largest eigenvalue λ_1 is excluded from our analyses because it is always zero in our method. The results indicate that using the sequence lengths 200, 500, and 2,000 sites had very similar performances when looking at how the mean squared errors of eigenvalues changed. If the evolutionary distances are small, the mean squared errors are very close to zero but when the evolutionary distances increase, the mean squared errors increase dramatically. This suggests 200 sites as the minimum sequence length which we use in following simulations. Figure 2.1 also indicates that using pairs with similarities of 0.25 or smaller substantially increases the mean squared errors. Our other results, which are not presented, show that pairs starting from similarities 0.25 or smaller have substantial probability of giving negative valued estimates for the expression within the logarithm in (2.2) or a value that is far from the true value. In the following simulations, we used 0.25 as a cutoff similarity, ignoring any pair with similarity lower than this.

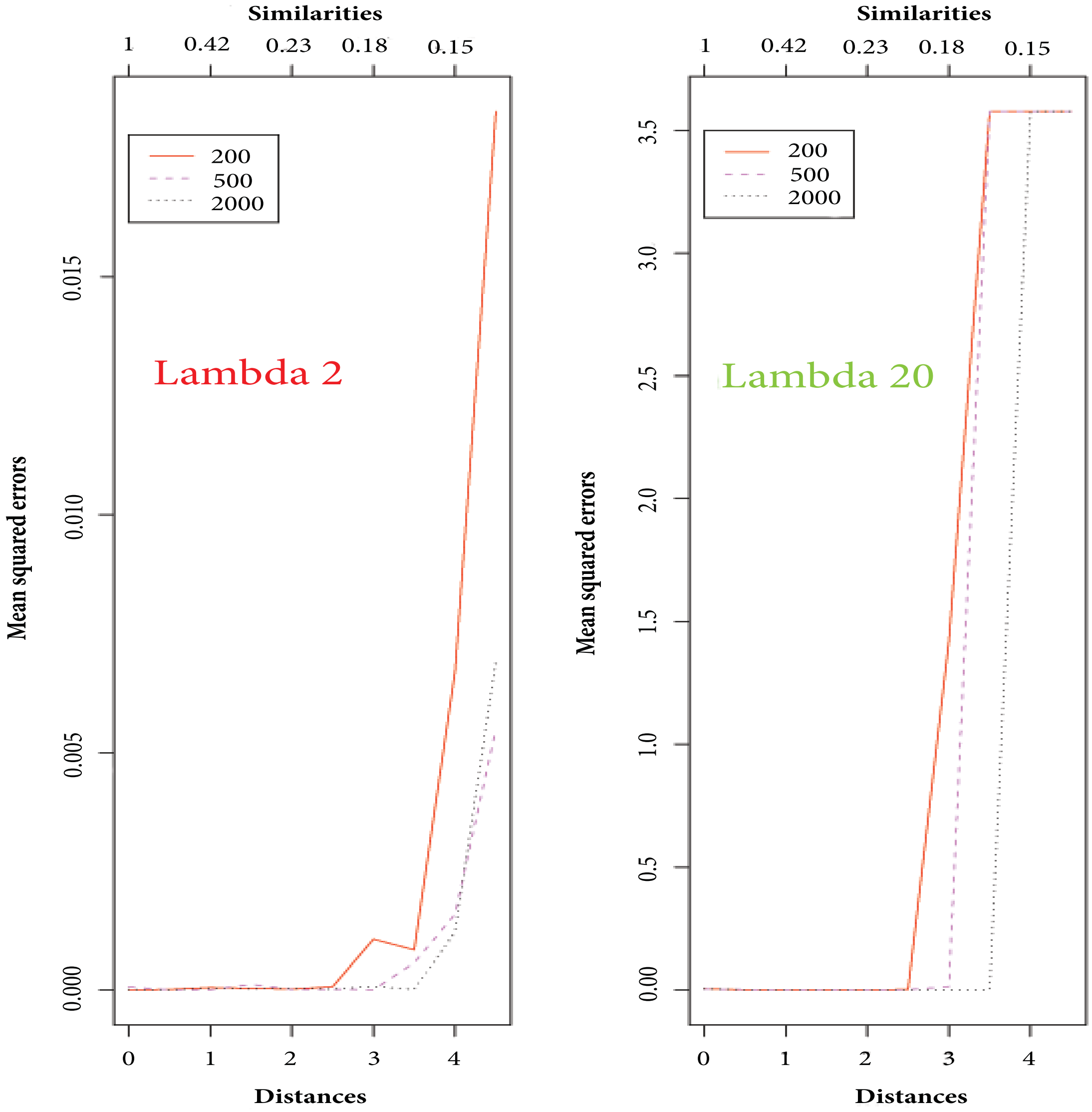


Figure 2.1: The mean squared errors of estimates of λ_2 and λ_{20} at 10 distances 0, 0.5, 1.0, 1.5, 2.0, 2.5, 3.0, 3.5, 4.0, and 4.5 using different sequence lengths (200, 500, 2,000 sites) in simulation 1.

The results of simulation 2 are shown in Table 2.1 and Figure 2.2 applying the binning strategy for data sets with 1,000, 5,000, and 10,000 pairs data sets. For each given binning strategy, the average MSE, averaged over the 19 eigen values and 100 simulations, was obtained and is referred to as aveMSE. We also obtained a measure of bias which we denote aveBias. For a given binning strategy and eigenvalue, we obtain an estimate of the bias as the average estimated eigenvalue, over 100 simulations, minus the generating eigenvalue. The average absolute value of these biases over eigenvalues gives aveBias. We also calculated aveMSEs and aveBiases for estimates of rate matrices. From Table 2.1 and Figure 2.2, we can see clearly that equal size strategy does better than equal width strategy both in terms of aveMSEs and aveBiases. Considering that the distribution of the probabilities of differences in the population is not uniform, the equal width bins may suffer from uneven number of pairs in each bin. For the bin with few pairs, the information obtained from this bin may not be enough to get an accurate estimate. Consequently we recommend using the equal size binning strategy which we do in the following simulations investigating the optimum number of bins.

Table 2.1 gives results for the equal size binning strategy for various choices of the number of bins with different data sizes. Over three different sizes of data sets, 1,000, 5,000, and 10,000 pairs, binning into three, four and five bins have smaller values of aveMSEs and aveBiases than other binning strategies with small differences among these three binning strategies. Overall, we can see that binning into four bins is a marginally better than three or five bins.

A valid rate matrix should have the off-diagonal elements greater than or equal to zero. Table 2.1 also gives the average numbers of negative entries of 380 off-diagonals over 100 simulations. For any sizes of data sets in Table 2.1, the smaller the numbers of bins, the fewer negative off-diagonals. We note that with

Table 2.1: The summary of aveMSEs, aveBiases of eigenvalues and rates, the negative off-diagonals, and relative changes of rates

			1,000 pairs			
bins	e.MSE	e.Bias	q.MSE	q.Bias	neg.diag	entries > 10%
1	1.69E-03	1.70E-03	7.82E-05	7.82E-05	0.84	217
2	1.30E-04	3.10E-04	2.65E-05	3.31E-05	3.52	57
3	4.07E-05	1.11E-04	2.76E-05	2.51E-05	4.32	37
4	4.05E-05	9.86E-05	3.11E-05	2.50E-05	4.92	36
5	5.33E-05	9.18E-05	3.54E-05	2.63E-05	5.34	41
6	6.27E-05	7.01E-05	3.64E-05	3.40E-05	5.36	44
7	7.01E-05	6.91E-05	3.94E-05	4.17E-05	5.54	43
8	1.08E-04	1.08E-04	5.75E-05	6.63E-05	6.56	48
			5,000 pairs			
bins	e.MSE	e.Bias	q.MSE	q.Bias	neg.diag	entries > 10%
1	1.79E-04	1.79E-04	6.68E-05	6.68E-05	0.0	213
2	1.60E-04	1.65E-04	7.33E-06	1.71E-05	0.24	54
3	5.11E-05	1.32E-04	6.05E-06	7.45E-06	0.38	22
4	4.02E-05	8.59E-05	7.46E-06	5.73E-06	0.42	17
5	4.57E-05	5.67E-05	8.89E-06	6.27E-06	0.46	16
6	5.37E-05	5.11E-05	1.04E-05	9.38E-06	0.48	21
7	6.18E-05	5.47E-05	1.19E-05	2.09E-05	0.50	31
8	7.16E-05	8.85E-05	1.40E-05	5.80E-05	0.60	38
			10,000 pairs			
bins	e.MSE	e.Bias	q.MSE	q.Bias	neg.diag	entries > 10%
1	1.78E-03	1.78E-03	6.49E-05	6.49E-05	0.0	214
2	1.60E-04	3.03E-04	5.21E-06	1.49E-05	0.2	43
3	5.19E-05	1.23E-04	3.69E-06	5.32E-06	0.08	11
4	4.13E-05	7.80E-05	4.89E-06	3.52E-06	0.08	10
5	4.65E-05	5.57E-05	6.18E-06	3.91E-06	0.10	12
6	5.39E-05	4.63E-05	7.36E-06	7.05E-06	0.10	14
7	6.14E-05	6.57E-05	8.71E-06	1.53E-05	0.10	21
8	7.02E-05	0.000121334	1.04E-05	4.65E-05	0.14	29

bins: number of bins used
e.MSE and e.Biase: the averaged aveMSEs and aveBiases of eigenvalues
q.MSE and q.Biase: the averaged aveMSEs and aveBiases of rate matrices
neg.diag: averaged negative diagonals among 380 entries among 100 runs
entries > 10%: the number of entries which have the relative errors greater than 10%

four bins, the average numbers of negative entries of 380 off-diagonals drops from 4.92 to 0.42 entries as we increase the number of pairs from 1,000 to 5,000 pairs. This is a strong indication that the data size 5,000 pairs is reasonable. In practice, we suggest replacing any negative estimates of off-diagonal elements by zeros and adjusting diagonal values accordingly.

To investigate biases of rate matrix estimation, we calculated the relative error by $abs((\text{estimate} - \text{jtt})/\text{jtt})$, where estimate is the means of estimates of rates over 100 simulations. The number of entries with relative error greater than 10% drops,

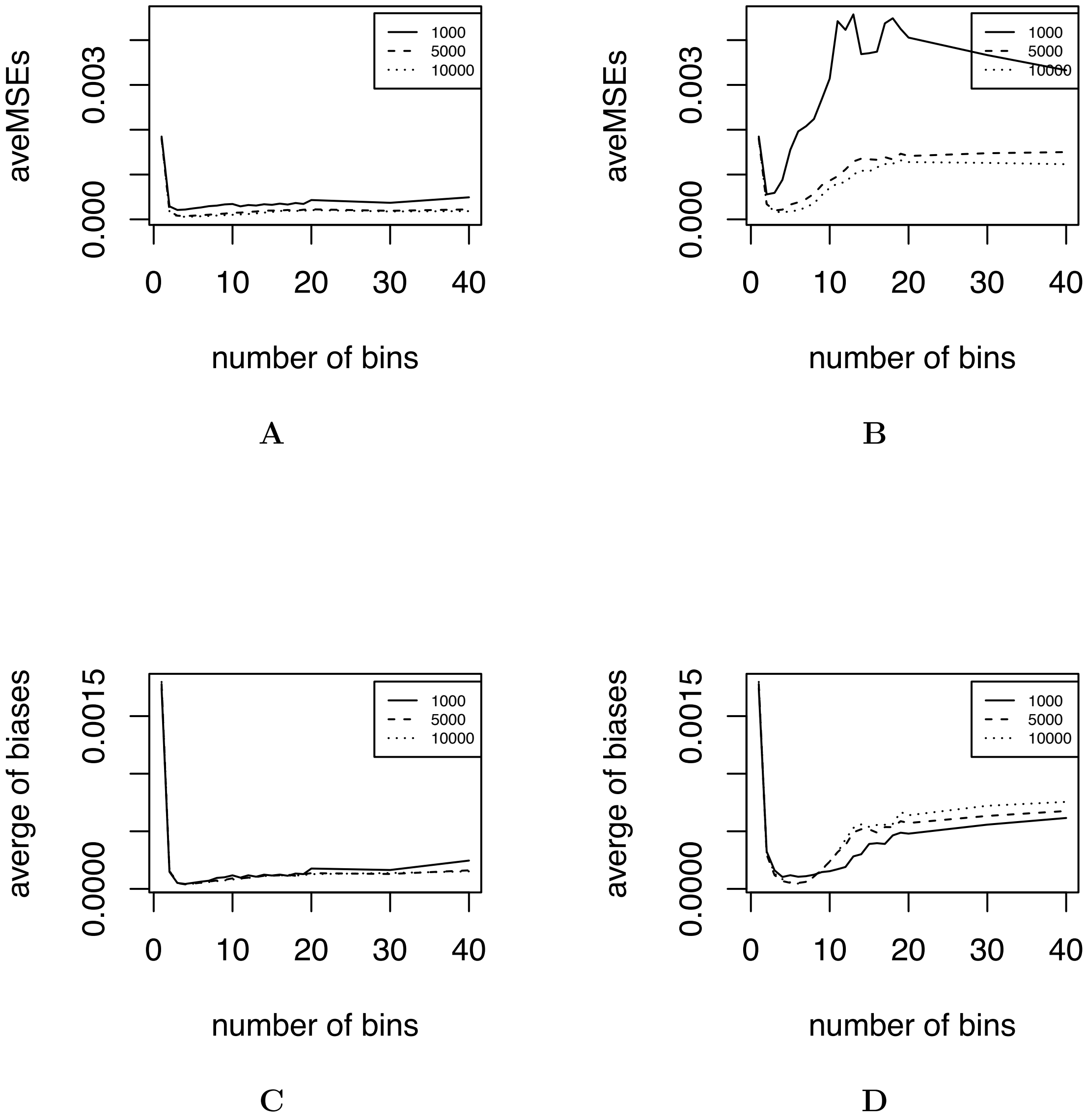


Figure 2.2: The overall average of means of squared errors and average of biases of the estimates of the eigenvalues in different binning strategies with sequence length 200 sites and 1000 to 10000 pairs of sequences for simulation 2. **A** and **C** in the column of left-hand side are the results of the equal size strategy; **B** and **D** in the column at right-hand side are for the equal width one. In each graph, the number of the bins are from 1 to 20 plus 30 and 40.

as expected, as the number of pairs increases, with the largest drop from 1,000 to 5,000 pairs. Together with the above results, this suggests reasonable estimates can be obtained with 5,000 pairs and sequence lengths of 200 sites using four bins.

The results of the simulations 3 and 4 indicate what we can be expected estimating a rate matrix from a small data set either with fixed sequence length or a fixed number of taxa. Figure 2.3 presents the results of simulation 3. In this simulation, the data sets had 20 taxa with sequence lengths from 200 to 10,000 sites. From Figure 2.3, we can see that both the aveMSE 0.918 and the aveBias 0.004 of the 200 sites are the worst. Then they drop to 0.0359 and 0.0026 with 500 sites. When increasing data set sizes from 500 sites, the aveBiases are roughly stable around 0.0026; aveMSEs change gradually to 0.004. For the average numbers of negative entries of 380 off-diagonals, it is not surprising to see 25.08, 11.4, 4.65, 1.63, and less than 0.88 entries for 200, 500, 1,000, 2,000, and 3,000 or more sites. Overall, it looks like that one can get a reasonable estimate of the rate matrix only with 500 $\sim$ 1000 sites for a 20 taxa data set.

In the simulation 4, the sequence length was fixed at 500 sites. We found that aveMSEs and the average numbers of the negative off-diagonal elements are very similar for 80 to 100 taxa. The aveBiases are close for 30 to 100 taxa. With 30 taxa, one can expect about four negative off-diagonal entries among 380 in an estimate of the rate matrix on average.

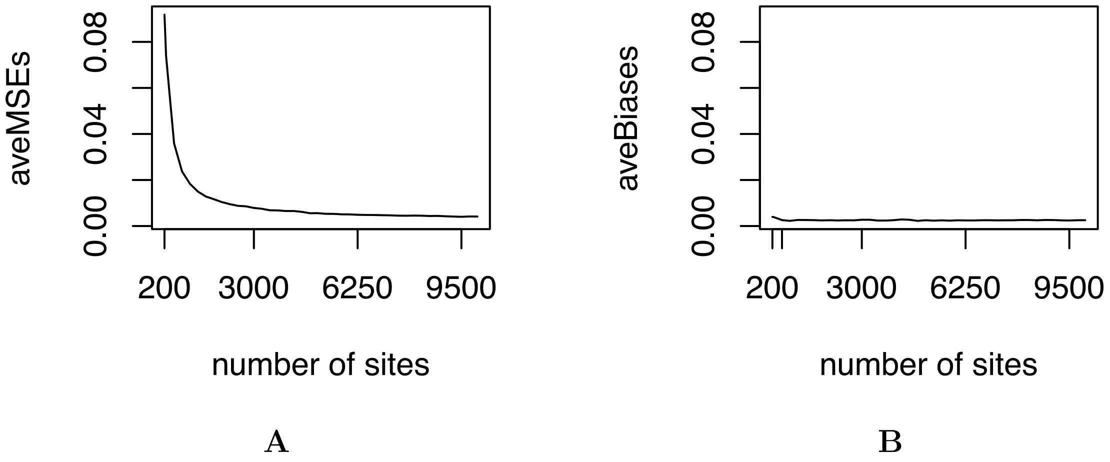


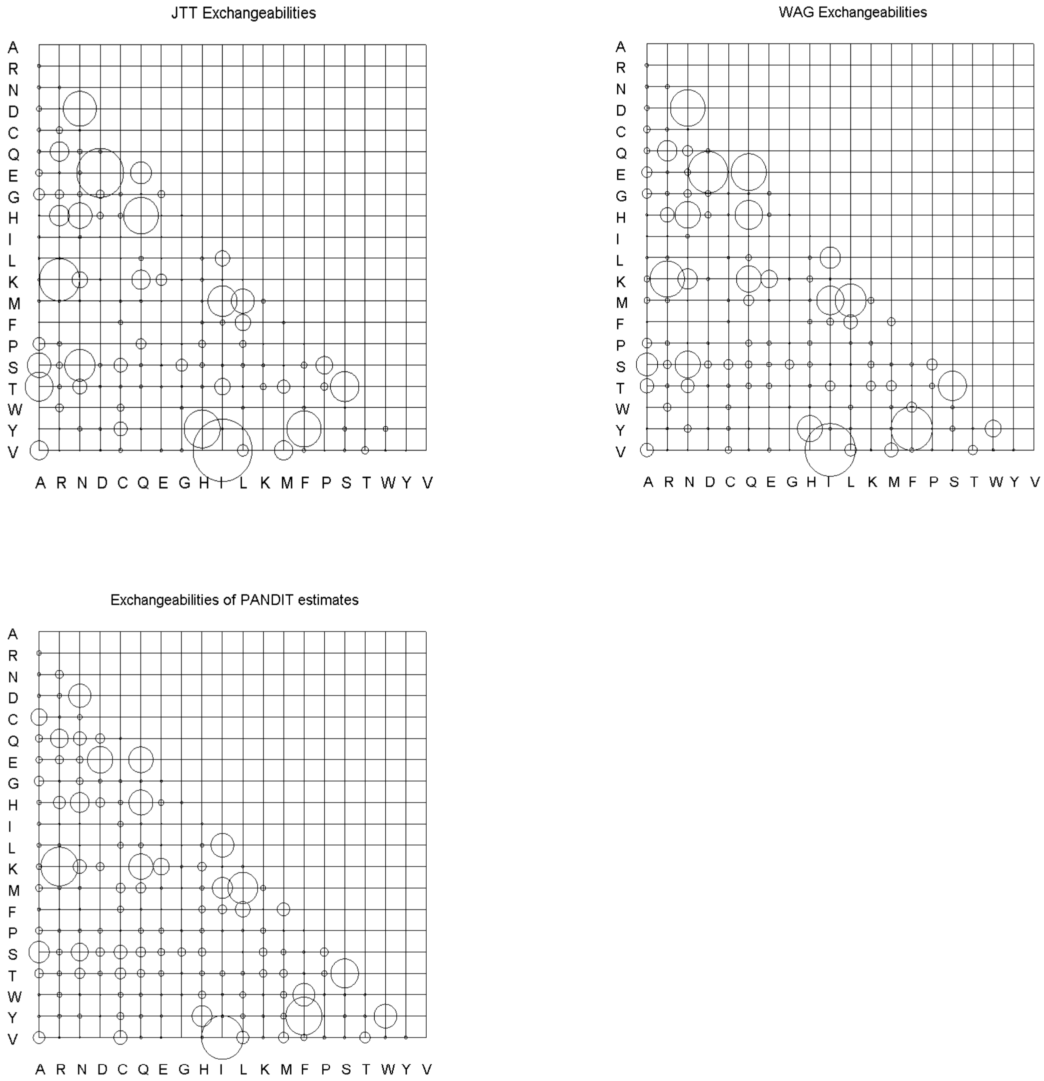
Figure 2.3: The overall averaged mean squared errors and averaged biases of the estimates of the Qs over different sequence lengths for simulation 3. **A** is the results for the aveMSEs, average of means of sums of squared errors, over 380 off-diagonal entries; **B** is the results for the aveBiases, average of biases, over 380 off-diagonal entries.

2.6 An Empirical Rate Matrix from the PANDIT Database

Using the current version of the PANDIT database of Whelan et al. (2006), we obtained an empirical rate matrix and compared it to those of the JTT and WAG models. Based on our simulation results, we used four equal-sized bins. Pairs were included in the estimation if their similarities were 0.25 or higher with 200 or more sites. With these criteria, a total of 183,844 pairs from 2,300 protein families were selected.

Our estimate of the instantaneous rate matrix has no negative off-diagonal entry. The sums over the rows are in the range $(-10^{-16}, 10^{-16})$. Hence, it is a valid estimate of the instantaneous rate matrix. In Figure 2.4, the exchangeabilities of the JTT, WAG and our estimates (labeled PANDIT) are presented. We can see that the three plots in Figure 2.4 are similar in pattern, but different in detail. As indicated in Jones et al. (1992), the pairs used to construct the JTT rate matrix had the 85% or larger similarities. By contrast, the WAG matrix was estimated using 3905 globular protein amino acid sequences in 182 protein families without 85% pairwise similarity constraint. For our estimate, we used a data set which has 2,300 protein families. Comparing our estimates with the JTT and WAG models, the bubbles of our estimates look smoother than the WAG ones which in turn are smoother than the JTT estimates. This reflects the fact that our procedure is able to use more data and more divergent sequences than the JTT and WAG models. It has the advantage that unlike WAG, we do not need an estimated tree.

The alignments used in our estimation are originally from the Swiss-Prot database (Boeckmann et al. 2003) aligned using ClustalW (Thompson et al. 1994). The PAM series matrices were used when obtaining scores for alignments. Pairs of different amino acids that are unlikely to arise as a consequence of substitution under a PAM model receive a relatively poor score. Consequently, methods that infer rates of substitution from such alignments will tend to see fewer pairs of character states that are unusual under the PAM model than they might have had a different scoring method been used. This may provide part of the explanation for the similarity between the JTT model and our estimates in the bubble plot.


JTT Exchangeabilities

WAG Exchangeabilities

Exchangeabilities of PANDIT estimates

Figure 2.4: Exchangeabilities of the JTT, the WAG and the PANDIT estimates

Conclusions

Our simulation results show that the binning strategy is an effective solution to overcome some of difficulties of the AB method when using amino acid data. Our simulation results indicate that the method works well without requiring phylogenetic reconstruction. While our main application was to a large database, our method can also be used to estimate a rate matrix from a relatively small data set.

The method requires some decisions about what bins should be used. Our simulation results suggest that an equal size binning strategy is preferable to an equal width strategy. In addition, pairs with similarities lower than 0.25 give rise to unstable estimation of eigenvalues and should be removed. Simulation results pertaining to the number of bins were less clear but it appears that fewer rather than more bins usually give rise to to better estimation. In the absence of additional information, our simulations suggest using four equal sized bins.

Chapter 3

Identifiability and the Barry Hartigan Model

3.1 Introduction

Phylogenetic estimation using model-based distance, maximum likelihood and Bayesian approaches have exploded in popularity in the last decade. Markov models employed by these approaches (e.g., the general time reversible (GTR) model) typically assume that nucleotide or amino acid sequences evolve through a homogeneous, stationary or reversible process over edges of the tree of Life (Felsenstein, 2004). However, in some cases, molecular sequence evolution will fail to satisfy these assumptions; nucleotide or amino acid frequencies can sometimes vary greatly between sequences under examination and the use of Markov models that fail to account for this property can positively mislead phylogenetic estimation (Foster and Hickey 1999, Galtier and Gouy 1995). Jayaswal et al. (2005) have showed that, for such cases, a more general Markov model proposed first by Barry and Hartigan (1987) (the 'BH' model) is a useful alternative to the GTR family of models because it makes many fewer assumptions, namely that: (3.1) evolution follows a Markov process across each edge and, (3.2) the data patterns at each site are independently and identically distributed. As a result, the BH model allows the character frequencies to differ across nodes in the tree and the evolutionary process to differ not only across edges but along the two directions on an edge. Here we show that despite the fact that the BH model has been shown to improve accuracy of phylogenetic estimation under some conditions, there are potentially serious problems with the identifiability of its parameters.

An identifiability problem arises under the BH model, because, as shown in detail below, different sets of parameters can lead to the same probability distribution of data patterns under the model and hence the same likelihood of the data. Specifically, we will show that there always exist multiple different sets of transition probabilities along edges and different sets of character base frequency vectors at internal nodes that will lead yield the same likelihood. If the BH model is being used to infer the tree topology alone, this will not create a problem. This follows from the main result of Steel (1994) which shows that LogDet distances are tree-additive for the BH model, implying that sets of pairwise distributions for different topologies are always different. However if researchers are interested in the nucleotide transition probabilities along edges of the phylogeny, the nucleotide frequencies at internal nodes, or approximations to edge lengths as, for example, proposed by Jayaswal et al. (2005), non-identifiability becomes a major problem. Below, we explain the nature of this non-identifiability and the effect it has on estimation of parameters.

3.2 The Identifiability Problem

Theorem 4.1 of Chang (1996) shows that if estimates are constrained to be within a class of transition matrices that is reconstructible from rows, those matrices are identifiable. The identifiability problem for the BH model arises because it places no such restrictions on the transition matrices that can be estimated. To illustrate the problem, assume that we have obtained three sequences from species a, b, and c in Figure 3.1, and the sequences are made up of strings of two states "0" and "1". Also assume that $P^a, P^b, \ and \ P^c$ are transition matrices of three branches (i, a), (i, b) and (i, c). For a site which has the states (x_a, x_b, x_c), the site pattern probability is

$$P(x_a, x_b, x_c) = \pi_0 * P^a_{0,x_a} * P^b_{0,x_b} * P^c_{0,x_c} + \pi_1 * P^a_{1,x_a} * P^b_{1,x_b} * P^c_{1,x_c} \tag{3.1}$$

where π_0 and π_1 are the base character frequencies at the internal node i. Now if the rows of $P^a, P^b, \ and \ P^c$, are permuted and π_0 and π_1 are exchanged, then we have $\pi^s_0 = \pi_1, \pi^s_1 = \pi_0$; $P^{sa}_{0,x_a} = P^a_{1,x_a}$, $P^{sb}_{0,x_b} = P^b_{1,x_b}$, $P^{sc}_{0,x_c} = P^c_{1,x_c}$ and etc. The pattern probability of this site after this permutation is shown in formula 3.2.

$$\begin{aligned} P^s(x_a, x_b, x_c) &= \pi^s_0 * P^{sa}_{0,x_a} * P^{sb}_{0,x_b} * P^{sc}_{0,x_c} + \pi^s_1 * P^{sa}_{1,x_a} * P^{sb}_{1,x_b} * P^{sc}_{1,x_c} \\ &= \pi_1 * P^a_{1,x_a} * P^b_{1,x_b} * P^c_{1,x_c} + \pi_0 * P^a_{0,x_a} * P^b_{0,x_b} * P^c_{0,x_c} \\ &= P(x_a, x_b, x_c) \end{aligned} \tag{3.2}$$

Equation 3.2 implies that the probabilities of data patterns with and without permuting the rows of the conditional probabilities along three edges and the base character frequencies at internal node i remain the same. Therefore this model is not identifiable. Note that the same permutation was applied to each of the transition matrices. More generally, for sequence data with m different states and three taxa, there are $m!$ permutations which have the same site pattern distribution. Furthermore, for any topology which has more than one internal node, the problem is compounded because multiple permutations of transition matrices can be performed at each internal node without a change in the likelihood of the data.

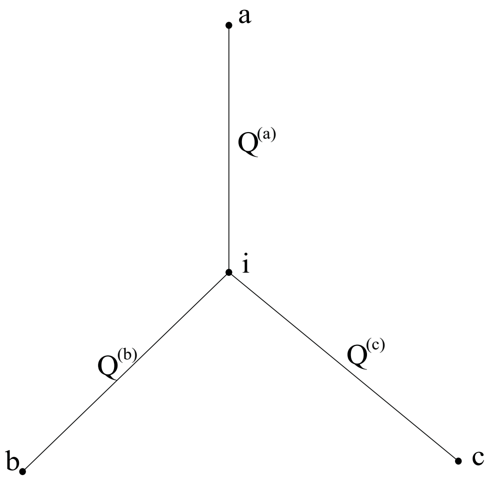


Figure 3.1: The three-taxon tree used in simulations. The $Q^{(e)}$ are oriented so that rows refer to the state at the internal node. The simulating model has the same $Q^{(e)}$ for each edge determined from a GTR model with edge length 0.5, stationary frequencies $0.1, 0.35, 0.4, 0.15$ and exchangeabilities $0.1, 1, 0.5, 0.5, 1$ and 0.1 for AC, AG, AT, CG, CT AND GT

.

3.3 Simulations

To demonstrate how and why the non-identifiability of the BH model may create problems, we simulated large sequence-length datasets under the widely-used GTR model, which is a special case of the BH model. For these simulated datasets, we know the true generating parameters and thus we can make a comparison between the true values and the estimates of parameters. With long sequence lengths, uncertainty due to estimation is minimal, and therefore large departures of estimates from the true values can be attributed to non-identifiability effects. For the simulation, we selected a very simple three-branch tree (Figure 3.1) and simulated data with the Seq-Gen program (Rambaut and Grassly, 1997) under a GTR model with arbitrarily chosen parameters: as a stationary nucleotide frequency vector of $\Pi = \{0.10, 0.35, 0.4, 0.15\}$, state exchangeabilities of $r = \{0.1(a \leftrightarrow c), 1(a \leftrightarrow g), 0.5(a \leftrightarrow t), 0.5(c \leftrightarrow g), 1(c \leftrightarrow t), 0.1(g \leftrightarrow t)\}$ and a sequence length of 100,000 sites. For this data set, we estimated parameters of the BH model using the software described in Jayaswal et al. (2005).

To determine whether our estimated matrices have the correct permutation relative to the model used for simulation, we employed a method that calculates the sum of squares between the entries of true joint probability matrices, $Q^{(e)}$, and the estimated matrices, $\hat{Q}^{(e)}$, as follows:

$$SS_{Qs} = \min_{s \in S} \sum_{e \in E} \sum_{k,j \in \{A,C,G,T\}} (Q^{(e)}_{k,j} - \hat{Q}^{(e)s}_{k,j})^2 \tag{3.3}$$

Here S is the set of 24 permutations of rows, E contains the three branches and

$\hat{Q}^{(e)s}_{i,j}$ is the s^{th} permutation of $\hat{Q}^{(e)}$. SS_{Qs} gives the minimum departure of estimates from the true matrices over all permutations of the rows of the estimated matrices; the same permutation is applied for each edge. If minimum sum of squares is attained without requiring permutation, the best-fitting permutation is the 'correct' permutation, otherwise it is a 'wrong' permutation. For the large sequence length datasets we considered, only one permutation of the matrices gave a small sum of squares.

3.4 Results and Discussion for Permutations of BH Estimates

Here we show the results for a simulated dataset with 100,000 sites. To avoid difficulties with local maxima and to investigate whether the correct permutation tends to get estimated in practice, we used 100 sets of randomly generated matrices along the three branches in Figure 3.1 as initial values for ML estimation and obtained 100 sets of estimates for the single dataset. Of these 100 sets of estimates, 97 converged on the global optimum with log-likelihoods within 0.5 of -349,111.0, whereas three of them were local maxima with log-likelihoods less than -349,366.7. Below we restrict attention to the 97 cases where the global optimum was estimated.

For these 97 estimates, SS_{Qs}s ranged from 0.000067 to 0.00018 indicating that the actual values of the entries of estimated matrices were very close to the entries in the true generating matrices once they were permuted appropriately. However, for 93 of the cases, the estimated matrices did not correspond to the original permutation of the matrix; the best-fitting permutations covered 23 of the 24 possible permutations (Table 3.1). Thus the non-identifiability problem with the BH model arises because the correct permutation of the estimated matrices cannot be recovered in practice.

Table 3.1: The numbers of times that each of the 24 permutations was the best-fitting permutation among the 97 runs

permutation	count	permutation	count	permutation	count
1234	4	2314	6	3412	4
1243	5	2341	6	3421	7
1324	5	2413	4	4123	4
1342	2	2431	1	4132	1
1423	2	3124	5	4213	3
1432	4	3142	3	4231	4
2134	4	3214	5	4312	0
2143	5	3241	7	4321	6

One of the potential benefits of the BH model is that it allows the researcher to both accommodate changing state (nucleotide or amino acid) frequencies during phylogenetic estimation as well as estimate the ancestral character state frequency vector at internal nodes of the tree. The latter parameters are of general interest to molecular evolutionists as they may give insights into the physical properties of ancestral molecules and environments (for example, see Boussau et al. 2008). Unfortunately, one result of the non-identifiability problem we describe is that the ancestral base character frequency vector of the internal node will also only be accurately estimated up to permutation. To illustrate this, we estimated the base character frequency vectors of internal node i for the five most frequent permutations recovered above (Table 3.2). Because incorrect permutations were estimated, the estimated base frequencies are drastically different from the true values with an averaged sum of square error of approximately 0.13 as compared with 0.002 for the correct permutations.

Table 3.2: The mean vectors of estimates of internal node frequency vectors for the five most frequent permutations

permutation	counts	mean vector of the estimates
3214	7	(0.15, 0.35, 0.11, 0.39)
3421	7	(0.15, 0.40, 0.11, 0.34)
2314	6	(0.40, 0.11, 0.34, 0.15)
2341	6	(0.15, 0.11, 0.34, 0.39)
4321	6	(0.15, 0.40, 0.34, 0.11)
true vector		(0.10, 0.35, 0.40, 0.15)

Another serious problem can occur if researchers wish to estimate edge lengths for their phylogenetic tree based on the estimated BH model parameters. For the BH model, the entries in the joint probability matrices along edges are optimized directly. No edge length estimation is involved as the evolutionary processes along edges need not correspond to a continuous time Markov process. However, Jayaswal et al. (2005) described a way to extract an approximate edge length estimate from the BH model parameters by assuming a continuous time Markov process. A difficulty arises because the method requires logarithms of all of the eigenvalues of the conditional probability matrices of the two evolutionary directions along a branch to exist. Unfortunately, we find that for most estimated permutations this is not the case. For example, for the eigen decompositions of the transition matrices along branches of 24 permutations of the example above we found that typically only one permutation has all positive eigenvalues whereas the others all have complex and/or negative valued eigenvalues. For idealized simulation cases where the generating model is GTR and there is sufficient data, the correct permutation of the BH estimate will correspond to a case where there the conditional probability matrices have positive eigenvalues, as in our example. However, for real data that have evolved under heterogeneous historical conditions there is no guarantee that any of the permutations of these matrices will meet this condition or, if they do, that they correspond to the correct permutations. For example, we fit the BH model to the *Plasmodium* species phylogenomic dataset and tree described by Davalos and Perkins (2008) which had eight taxa and 77313 sites after gaps were removed. We estimated the parameters using the BH model with random initial starting values. For each of the 13 edges of the tree in Figure 1(A) of Davalos and Perkins (2008), we found that the transition matrices always had negative eigenvalues even after ignoring the complex-valued portions of the eigenvalues. Thus, the approximate method of Jayaswal et al. (2005) could not be used to obtain edge lengths for this case.

From the foregoing discussion it should be clear that, for BH model parameter estimates to be useful to researchers, a method for estimating the correct permutation is necessary. One possibility is given in the implementation of Jayaswal et al. (2005), where they recommended initial parameter values for BH optimization be $\frac{1}{8}$ on the diagonal and $\frac{1}{24}$ on the off-diagonal elements. This satisfies the condition of ‘diagonal largest in column’ (DLC) for transition matrices in both evolutionary directions. The DLC condition is discussed in Chang (1996) as a potential condition for identifiability. If the true evolutionary process yields transition matrices with this property, then starting the optimization with initial parameters that satisfy DLC could help with the estimation of the correct permutation. We did an experiment randomly generating 100 datasets each with 1000 sites under our generating GTR model which does satisfy the DLC condition. For 100 sets of estimates of transition probability matrices obtained using Jayaswal and colleagues' recommended initial values, we always obtained the correct permutation.

However, in practice part of the reason for considering a BH model is to fit nonstationary processes; in the latter, there is no guarantee that the DLC condition will hold for the correct permutation. To demonstrate this, we simulated under a nonstationary model using the tree in Figure 3.1. For this generating model, the frequency vector was $\{0.45, 0.05, 0.05, 0.45\}$ for a GTR model along edge (a, i); the frequency vector was $\{0.05, 0.45, 0.45, 0.05\}$ for edges (i, b) and (i, c). All three edges shared the same exchangeability vector {1.0, 5.6, 1.0, 1.0, 5.6, 1.0} for $a \leftrightarrow c, a \leftrightarrow g, a \leftrightarrow t, c \leftrightarrow g, c \leftrightarrow t, g \leftrightarrow t$ and the root frequency vector is $\{0.1, 0.4, 0.4, 0.1\}$. The model was nonstationary because different GTR models were used on the other edges. Here, although the transition matrices do satisfy the DLC condition, the joint probability matrices do not. In this experiment, the estimates obtained based on optimization from Jayaswal and colleagues' recommended initial parameter settings for the joint probability matrices did not correspond to the correct permutations and the eigenvalue vectors of the estimated transition matrices along the three edges had negative values.

3.5 A Parsimony Method of Estimating Permutations

Below we describe a simple approximate method for finding the correct permutations that may be useful in practice and illustrate it with reference to the topology of Figure 3.1. If we assume that the evolutionary processes along the three branches are relatively close to stationarity, then it is likely that the ancestral nucleotide frequency vector of internal node i will be relatively similar to the frequencies of the two nearest nodes. For instance, if the average of the character base frequency vectors of leaf nodes a and b in Figure 3.1 were $\{0.85, 0.15, 0, 0\}$, given the edge lengths shown, the correct permutation of the estimate the internal node i nucleotide frequency vector should be closer to $\{0.85, 0.15, 0, 0\}$ than other possible permutations (e.g., $\{0.15, 0.85, 0, 0\}$). Based on this intuition, we propose a least squares method for estimating the correct permutation. Let $\hat{\Pi}_a, \hat{\Pi}_b, \hat{\Pi}_c,\ \mathit{and}\ \hat{\Pi}_i$ denote the estimated frequency vectors at the nodes a, b, c, and i. For any permutation s of estimates, we compute $||\hat{\Pi}_i^s - (\hat{\Pi}_a + \hat{\Pi}_b + \hat{\Pi}_c)/3||^2$, where $a, b,$ and c are nearest nodes of internal node i in subtree (a, b, c, i). The best-fitting permutation is computed by the following criterion.

$$SS_{\pi s} = \min_{s \in S} ||\hat{\Pi}_i^s - (\hat{\Pi}_a + \hat{\Pi}_b + \hat{\Pi}_c)/3||^2 \tag{3.4}$$

Testing our idea, we used the estimates obtained from our previous simulation experiments, and estimated the permutation using (3.4). We found that in all cases examined the sums of squares of the best-fitting permutations were smaller than others. Secondly, we checked whether the s obtained by this strategy was the true correct permutation and verified that for the cases where the estimation converged on global optimal likelihood, the correct permutation was always selected by our method. In general, we expect that this method will perform well for real data sets where the nucleotide composition of sequences is changing gradually (rather than abruptly) over the tree topology.

To extend our proposal to more than three taxa, one could minimize the sum of (3.4) over all internal nodes. One approach to doing so is as follows. Given an ordering of the internal nodes, successively to determine the permutations minimizing (3.4) for each of the internal nodes. For each internal node, when minimizing, hold the frequency vectors at all other internal nodes fixed. Iterate the process until no further improvement in the sum is possible. Iteration is required here since the best permutation for node i depends on the frequencies at a. Consequently, if the permutation at node a changes when it is considered, the permutation for node i minimizing (3.4) might change as well.

Although the phylogenetic tree estimated by the BH model is an identifiable parameter, we have shown that there is a problem with identifiability of the joint probability matrix parameters. If ignored, this non-identifiability problem can mislead researchers interested in the ancestral character state compositions estimated by the BH method, or approximate edge lengths generated by existing implementations. We have proposed a solution that will select the correct permutation when the data evolves in a ‘close to stationary’ manner. However, if this assumption is not correct, then researchers should be aware of the potential pitfalls stemming from the identifiability problem we have discussed.

Chapter 4

Fitting Nonstationary General-time-reversible Models to Obtain Edge Lengths and Frequencies for the Barry-Hartigan Model

4.1 Introduction

The general Markov phylogenetic model first introduced by Barry and Hartigan (1987), known as the BH model, is very flexible. As with most models, it assumes an independent and identical distribution among sites but differs in that it allows separate discrete-time Markov processes to occur along edges. Work by Jayaswal et al. (2005) and Oscamou et al. (2008) have shown that this model has better phylogenetic estimation properties than simple models when evolutionary processes are nonstationary. However, we previously showed that the estimates of the BH model suffer from identifiability problems that lead to difficulties with correctly estimating internal node frequencies (Zou et al. 2011a). A further complication is that the BH model does not directly involve edge-length estimation and thus neither conclusions nor subsequent analyses (e.g., molecular clock estimation) can be made that require such information.

A model can be described as non-identifiable when is defined as cases where two or more distinct parameter settings yield the same probability distribution on the data. The maximum likelihood (ML) estimates of a BH model are joint probability matrices. When using the BH model to reanalyze a recently published multigene dataset from the malaria parasites of the genus *Plasmodium* described in Davalos and Perkins (2008), we found that permuting the rows of some of the joint probability matrices gave exactly the same likelihood. Lemma 4.1 of Chang (1996) states that in a three-taxon tree, if the conditional probability matrix is reconstructible from rows, the full model is identifiable. This restriction does not hold for the BH model and our observations from the *Plasmodium* dataset confirmed that the estimates of transition matrices of the BH model are not unique; there are always at least 24 ML estimates. This differs significantly from the general time reversible (GTR) model because, as shown by Allman et al. (2008), this model with four states is identifiable for all parameters. For the BH model, although the estimates of leaf node state frequencies match the observed frequencies for the corresponding taxa (Jayaswal et al. 2005), there is no guarantee that the estimates of nucleotide frequencies at the internal nodes will be fitted to the right permutations. In this paper, we define the best-fitting nonstationary GTR models along edges, referred to as nonstationary best-fit GTR (NSGTR) models, and propose a method that employs the sum of squared differences between NSGTR estimates and BH estimates to identify the best-fitting permutations.

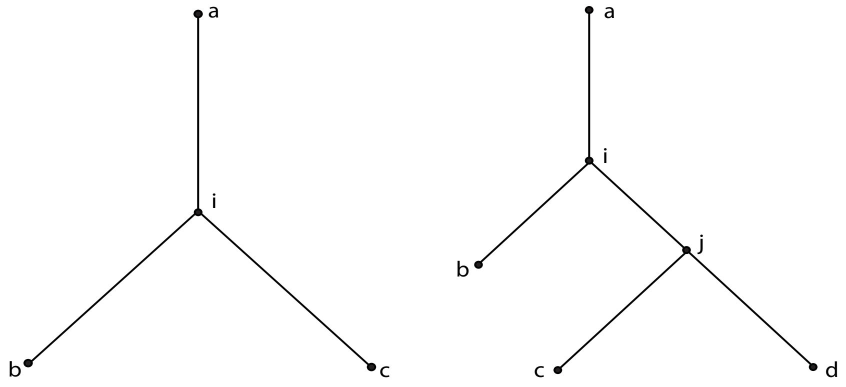
The general time reversible (GTR) Markov model assumes a continuous-time stationary process of nucleotide substitution occurs over the tree and is often used in phylogenetic estimation (see Chapter 13 in Felsenstein 2004). For this model, the evolutionary processes are at equilibrium and therefore the root position is not important as the entire topology shares one stationary frequency vector and one substitution rate matrix. The equilibrium assumption is restrictive and violations of this assumption for real datasets have been demonstrated by a number of studies including Yang and Roberts (1995), Foster and Hickey (1999), Foster (2004) and Ababneh et al. (2006). Two commonly used nonstationary models have been proposed in Yang and Roberts (1995) and Galtier and Gouy (1998). Yang and Robert's (YR) model used the Hasegawa-Kishino-Yano (HKY) model proposed in Hasegawa et al. (1985) as the base model but allowed each edge to have its own edge length and stationary frequency vector. However, in this case the entire tree still shares the same transition and transversion ratio. Galtier and Gouy (1998) suggested a simpler model (the GG model) that allows G+C-content to change throughout the tree. It uses the model in Tamura (1992) as a base and assumes a common transition and transversion ratio along all edges. This model is a special case of the YR model; for each edge, the GG model has 2 parameters less than the YR model. Compared to these two nonstationary models, the BH model is much more flexible.

While edge lengths were not the primary focus of Jayaswal et al. (2005), ideas were provided for estimating them. In most stationary models, edge lengths are interpretable as the expected numbers of substitutions. This interpretation is desirable for nonstationary models as well but does not apply for some current implementations. Jayaswal et al. (2005) estimated approximate edge lengths for the BH model by averaging the GTR distances for the two opposing evolutionary directions. However, this solution is difficult to justify if the process is nonstationary. Yang and Roberts (1995) pointed out that for nonstationary evolutionary processes, the base frequency vector of an edge will often change along the lineages. The edge lengths in Yang and Roberts (1995) were the conventional edge length t_s in $P(t_s) = e^{Rt_s}$ for an edge where $-\sum_j \pi_j R_{jj} = 1$, π_j is the stationary base frequency and R is the rate matrix; the edge lengths in the GG model were computed using the formula in the appendix of Galtier and Gouy (1998). Because both of these models are nonstationary, the edge length parameters they employ do not correspond to the conventional interpretation as the expected number of substitutions per site. Fortunately, Minin and Suchard (2008b) recently introduced a method to compute the conditional expected number of substitutions in the interval [0, t). This method can be used to obtain edge lengths interpretable as the expected numbers of substitutions per site for the YR, GG and NSGTR models.

4.2 Methods

The BH model assumes that all sites have the same distribution, that the evolutionary processes at all sites are independent and that a Markov process of substitution occurs along each edge. This process, however, need not correspond to a continuous-time Markov process. What is required is only that substitutions along an edge (a, b) occur with probabilities given by a matrix $P_{a,b}$ with positive entries and row sums equal to 1. Conditional upon the character state at an internal node, the process along adjacent edges is independent. For instance, for the star tree in Figure 4.1a, the processes along edges (a, i), (i, b) and (i, c) are independent of each other given the character state at i. The parameters of the joint probability matrices along the edges, may or may not correspond to a continuous Markov chain, but they must satisfy the internal consistency constraint for edges (a, i), (i, b) and (i, c) connected to internal node i whereby state frequencies at node i are the same regardless of the edge matrices. Because of this internal consistency constraint, the likelihood of the BH model does not depend on the position of the root of the tree.

In contrast, the NSGTR model described below is a nonstationary model that requires a root for its specification. The model assumes continuous-time GTR substitution processes along edges away from the root node. However, these GTR substitution processes are allowed to be different for different edges. In addition, the frequencies at the root node need not be the stationary frequencies of a GTR model for an edge connected to the root. For the k^{th} permutation $P^k_{a,b}$ of $P_{a,b}$, a NSGTR model $R^k_{a,b}$, $\Pi^{ks}_{a,b}$ and $t^k_{a,b}$ will be fitted and $P^{nk}_{a,b}$ will be recovered from this



a. Three taxa

b. four taxa

Figure 4.1: Trees of three and four taxa

NSGTR model. Define $SS^k_{ab} = \sum_{i,j}(P^k_{a,b}(i,j) - P^{nk}_{a,b}(i,j))^2$ as the sum of squares of the differences between $P^k_{a,b}$ and $P^{nk}_{a,b}$. The sum of squares, SS^k_{ba} is also computed for the reverse direction using the same procedure. After obtaining all SS^ks along all edges for both forward and reverse evolutionary directions, we sum over edges to get a total SS. As discussed above, this requires a root node since the evolutionary direction must be specified and, in absence of knowledge of position of the root, we consider all possible root nodes. In our approach, the internal consistency requirement plays an important role in determining the permutations. For instance, considering the three-taxon tree in Figure 4.1a, internal consistency gives $F^T_{a,i}\mathbf{1} = F_{i,b}\mathbf{1} = F_{i,c}\mathbf{1}$, where $F_{a,i}, F_{i,b}$, and $F_{i,c}$ are the joint probability matrices along edges of (a,i), (i,b) and (i,c); $\mathbf{1}$ is a column vector with ones as its elements. If the row permutation of $F_{i,b}$ of edge (i,b) is changed, the row permutation of $F_{i,c}$ and the column permutation of $F_{a,i}$ should be changed accordingly to satisfy internal consistency.

4.2.1 The Best Fitting Nonstationary GTR (NSGTR) Model

Under the assumptions of the NSGTR model above, for any edge (a, b), using the joint probability matrix $F_{a,b}$ and the frequency vector $\Pi_a = \sum_b F_{a,b}$, we can compute a transition matrix $P_{a,b} = \Pi_a^{-1} F_{a,b} = e^{R_{a,b} t_{a,b}}$, where $R_{a,b}$ is the instantaneous rate matrix of $P_{a,b}$; $t_{a,b}$ is the evolutionary time associated with $R_{a,b}$ of $P_{a,b}$. In Markov chain theory, if the process is at equilibrium, the largest of the eigenvalues is equal to 1 and the corresponding left eigenvector is the stationary frequency vector: $\Pi^s_{a,b} P_{a,b} = \Pi^s_{a,b}$, where $\Pi^s_{a,b}$ is the base frequency vector at the equilibrium. Using $P_{a,b} = e^{R_{a,b} t_{a,b}}$, we can compute $R_{a,b} t_{a,b}$. The GTR model obtained above is the best fit GTR model along edge (a, b) in the direction $a \rightarrow b$. Although there is only one correct direction, it is possible that there will be an $R_{b,a}$ and $t_{b,a}$ corresponding to a model in the reverse evolutionary direction that gives the same joint probabilities and, a priori, we may not know which is the correct direction. However, in general, $R_{a,b} \neq R_{b,a}$ and $t_{a,b} \neq t_{b,a}$ so knowing the direction of evolution matters and hence specification of a root of the tree will ultimately be necessary (discussed below in more detail).

In any case, when estimating the best fitting NSGTR model in direction $a \rightarrow b$, we first obtain $\Pi^s_{a,b}$ through eigenvector decomposition of $P_{a,b} = \Pi_a^{-1} F_{a,b}$ where $F_{a,b}$ and Π_a come from the BH model. Then we calculate a symmetric joint probability matrix $F^s_{a,b} = (\Pi^s_{a,b} \Pi_a^{-1} F_{a,b} + (\Pi^s_{a,b} \Pi_a^{-1} F_{a,b})^T)/2$. We calculate a symmetric estimate since this is implied by the corresponding true joint probability matrix for a time reversible model. Using $F^s_{a,b}$ and $\Pi^s_{a,b}$, we can then compute the rate matrix and the conventional edge length of the transition matrix $P^s_{a,b} = (\Pi^s_{a,b})^{-1} F^s_{a,b}$ for this edge. The rate matrix and conventional edge length are then estimated as log $P^s_{a,b}$ through eigenvector decomposition. In some cases, the eigenvector decomposition of $P^s_{a,b}$ gives negative eigenvalues and in these cases, an estimate of the rate matrix can not be obtained. However the best-fitting transition matrix that corresponds to a limiting case of the NSGTR model can still be obtained by setting the negative eigenvalues of $P^s_{a,b}$ to zero.

4.2.2 Iteratively Estimating the Permutations of Frequencies at Internal Nodes

As we have previously shown in Zou et al. (2011a), different permutations of the rows of BH substitution model can yield the same distribution of observed data. To illustrate, assume that (x_a, x_b, x_c) are the character states for species a, b, and c in Figure 4.1a and that these are each 0 or 1. Assume that π_0 and π_1 are the base character frequencies at the internal node i and that $P_{0,x_a}, P_{0,x_b}, P_{0,x_c}, P_{1,x_a}, P_{1,x_b},$ and P_{1,x_c} are the elements of transition matrices $P_{i,a}, P_{i,b},$ *and* $P_{i,c}$. Now if the rows of $P_{i,a}, P_{i,b},$ *and* $P_{i,c}$, are permuted and π_0 and π_1 are exchanged, then we have $\pi^s_0 = \pi_1, \pi^s_1 = \pi_0$; $P^s_{0,x_a} = P_{1,x_a}$, $P^s_{0,x_b} = P_{1,x_b}$, $P^s_{0,x_c} = P_{1,x_c}$, etc. Comparing the probabilities of the observed character states we obtain

$$\begin{aligned} P^s(x_a, x_b, x_c) &= \pi^s_0 * P^s_{0,x_a} * P^s_{0,x_b} * P^s_{0,x_c} + \pi^s_1 * P^s_{1,x_a} * P^s_{1,x_b} * P^s_{1,x_c} \\ &= \pi_1 * P_{1,x_a} * P_{1,x_b} * P_{1,x_c} + \pi_0 * P_{0,x_a} * P_{0,x_b} * P_{0,x_c} \\ &= P(x_a, x_b, x_c) \qquad (4.1) \end{aligned}$$

The pattern probabilities are the same before and after permuting regardless of the observed character states. Because of this lack of identifiability due to permutation, in a DNA dataset, 24 sets of estimates of joint probability matrices could give the same likelihood in the three-taxon tree of Figure 4.1a. For an internal edge which connects two internal nodes, there are 576 permutations of the rows and columns of the joint probability matrix for that edge that will give the same probability of data at two internal nodes; for an edge which is connected to an internal node and a leaf node or the root node, there are 24 permutations of the joint probability matrix for that edge that will give the same probability of data at the internal node.

We introduced a parsimony-like method for estimating the correct permutation of frequencies at internal nodes in Zou et al. (2011a). For this method to work the frequency vectors at two adjacent nodes should not be too different. Although we have used this method in the analysis of nonstationary data, it will be valuable to have a method that applies in more complicated and general cases. In the following, we introduce a method which can be used to estimate permutations of internal node frequencies without requiring frequencies at adjacent nodes to be similar.

The method for estimating permutations of internal node frequencies obtains the best-fitting NSGTR model above for each permutation of the BH estimated joint probability matrices. The estimated permutations along all edges are taken as those that give the overall minimum distances between the BH transition matrices and the NSGTR transition matrices as measured by the sum of squares of differences between these two matrices.

Using the four-taxon tree shown in Figure 4.1b, we will illustrate in more detail how the permutations are estimated. Our procedure takes all nodes in the tree as valid rooting positions and examines them one by one. For each rooting position we assign initial permutations for the joint probability matrices of all edges. Taking node a as the root, we compute the total minimum sum of squares $minSS$ from the tips of the tree to the root a. We start from the node j. In this subtree, we first fix the row permutation of the joint probability matrix of edge (i, j) and pick a permutation k_j which gives us $minSS_j$:

$$minSS_j^1 = \min_{k_j} \; SS_{ij}^{k_j} + SS_{jc}^{k_j} + SS_{jd}^{k_j} \tag{4.2}$$

The index k_j is the column permutation of F_{ij} and the row permutation of F_{jb} and F_{jc}. Having determined the permutation for j, we move to the node i. Keeping the column permutation of F_{ij} as k_j, we determine k_i giving $minSS_i$ using the same criterion of equation 4.2 but applied to the three edges connected to the node i. We calculate the total sum of squares of the first iteration using the SSs obtained for edges (a, i), (i, b), (j, c), (j, d) and (i, j), giving a decomposition of the sum of squares as $minSS^1 = SS_{ai}^{k_i} + SS_{ib}^{k_i} + SS_{jc}^{k_j} + SS_{jd}^{k_j} + SS_{ij}^{k_i,k_j}$. For the second iteration, k_i is the initial row permutation of F_{ij}. With this different initial permutation we repeat the process to obtain another set of k_i and k_j and $minSS^2$. Iterations continue as long as $minSS^m < minSS^{m-1}$. The permutation indices k_i and k_j of the final $minSS_{\mathcal{T}}$ are the estimated permutation of the BH estimates when rooting at node a. For each node in the tree, we repeat the procedure with that node being the root. The

k_i and k_j that minimize SS over all root choices are the estimated permutations of BH estimates.

In our experiments, for a given BH transition matrix, there was always only one permutation for which a GTR model with a set of valid instantaneous rate matrix and stationary frequency vector could be embedded. For permutations that do not give valid rate matrices, some corrections were needed. If $P^k_{a,b}$ has eigen-decomposition $U\Lambda U^{-1}$, then, up to a constant of proportionality, the rate matrix is estimated as $U \log[\Lambda] U^{-1}$. In practice, it is possible that some of the eigenvalues in Λ will be negative, making it impossible to take logarithms. In this case, we set the corresponding entries of $\log[\Lambda]$ to a large (in magnitude) negative number. One further correction was to set negative entries of the estimated rate matrix to 0 and then adjust diagonal entries accordingly so that rows of the rate matrix sum to 0.

4.3 Defining the Number of Substitutions

For most phylogenetic models, rate matrices are conventionally rescaled so that edge lengths are interpretable as expected numbers of substitutions. For a stationary model, if R is the rate matrix and Π the stationary matrix, R is rescaled so that $-\sum_j \pi_j R_{jj} = 1$. For nonstationary models, however, this rescaling will not necessarily give edge lengths with the correct interpretation (Minin and Suchard 2008b). Below we give a formula for the expected number of substitutions under our nonstationary model with an unscaled rate matrix.

Let $N(t)$ be the number of substitutions over an edge of length t; let R be the instantaneous rate matrix of a continuous time Markov model; $P(t) = e^{Rt}$ is the corresponding transition matrix. Let X_t denote the character state of the process at time t and let $R_{\mathcal{L}}$ denote the rate matrix R but with diagonal entries set to zero. Equation 2.3 of Minin and Suchard (2008b) gives

$$\mathbb{E}[N(t)I\{X_t = j\}|X_0 = i] = \int_0^t (e^{Rz}R_{\mathcal{L}}e^{R(t-z)})_{ij}dz$$

Thus, if $\beta_i = P(X_0 = i)$,

$$\begin{aligned}
\mathbb{E}[N(t)] &= \sum_{i,j} \beta_i \mathbb{E}[N(t)I\{X_t = j\}|X_0 = i] \\
&= \sum_{i,j} \beta_i \int_0^t (e^{Rz}R_{\mathcal{L}}e^{R(t-z)})_{ij}dz \\
&= \sum_{i} \beta_i \int_0^t \sum_{l,k\neq l} [e^{Rz}]_{il} \sum_{k\neq l} [R_{\mathcal{L}}]_{lk} \sum_{j} [e^{R(t-z)}]_{kj}dz
\end{aligned}$$

Since $\sum_j [e^{R(t-z)}]_{kj} = \sum_j P_{k,j}(t-z) = 1$ and $\sum_{k\neq l} R_{lk} = -R_{ll}$, we obtain that

$$\mathbb{E}[N(t)] = -\sum_i \beta_i \int_0^t \sum_l [P(z)]_{il} R_{ll} dz$$

For a time-reversible model, $P(t)$ has an eigenvector decomposition as $P(t) = Ue^{\Lambda t}U^{-1}$ where $e^{\Lambda t}$ is a diagonal matrix with i^{th} diagonal entry $e^{\lambda_i t}$; the i^{th} column of U gives the i^{th} eigenvector of P and $e^{\lambda_i t}$ gives the i^{th} eigenvalue; one of the λ_i is zero and the rest are negative. Using the eigenvector decomposition

$$\mathbb{E}[N(t)] = -\sum_{ijl} \beta_i R_{ll} [U]_{ij} [\int_0^t e^{\lambda_j z} dz][U^{-1}]_{jl} \tag{4.3}$$

When $j = 1$, the eigenvalue is zero which is the largest eigenvalue of the rate matrix. Thus $\int_0^t e^{\lambda_1 z} dz = t$. When $j \neq 1$, $\int_0^t e^{\lambda_j z} dz = \frac{e^{\lambda_j t} - 1}{\lambda_j}$. Thus,

$$\mathbb{E}[N(t)] = -t \sum_{il} \beta_i R_{ll} [U]_{i1} [U^{-1}]_{1l} - \sum_{il, j\neq 1} \beta_i R_{ll} [U]_{ij} [\frac{e^{\lambda_j t} - 1}{\lambda_j}][U^{-1}]_{jl} \tag{4.4}$$

Given an edge with transition matrix $P(t) = e^{Rt}$ for an unscaled R, if we take our edge length as $t_e = \mathbb{E}[N(t)]$ and rescale R by $\frac{1}{t_e} R_{ij}$, then t_e will be interpretable as the expected number of substitutions.

If the β_i are the stationary frequencies π_i for R, this gives the conventional rescaling where $-\sum_j \pi_j R_{jj} = 1$. The edge lengths coming from this rescaling will be denoted t_s and $t_s = t_e$ only in the case of a stationary model. For the NSGTR model the appropriate β_i required to calculate t_e are the frequencies at the starting node which need not coincide with the stationary frequencies.

The parameters of the BH model can be used to estimate a transition matrix P for any edge. In practice, to obtain t_e, we use the estimates of P^n instead of the estimates of P and substitute in (4.4).

4.4 Results and Discussion

4.4.1 Simulation Settings

We used the INDELible sequence simulator (Fletcher and Yang 2009) to create datasets to test the performance of our methods. The parameters of interest are joint probability matrices along edges, frequencies at all nodes and edge lengths. In our simulations, we evaluated both 'mild' and 'extreme' settings for parameters. The mild parameters corresponded to the parameter estimates obtained for the NS-GTR model fitted to a real phylogenomic dataset consisting of data from the genus *Plasmodium* (Davalos and Perkins 2008). The extreme parameters had extreme stationary frequencies and exchangeabilities but the same NSGTR edge lengths as in the mild parameters dataset. The tree in figure 1 (a) in Davalos and Perkins (2008) was treated as the true tree and has been reproduced in Figure 4.2.

4.4.2 Estimated Permutations

Our experience suggests that optimization of parameters under the BH model can sometimes yield local maxima. To check whether this was the case in our analysis, for a given dataset, simulated using the mild parameters setting, we randomly generated 100 sets of joint probability matrices under the true tree as our initial values to seed the optimization. When estimating parameters of the BH model for each of the 100 sets of joint probability matrices, we found that 94 out of the 100 had the same maximum log-likelihood up to one decimal place, -335974.1. In the following, we will ignore the 6 sets of estimates which had much smaller suboptimal

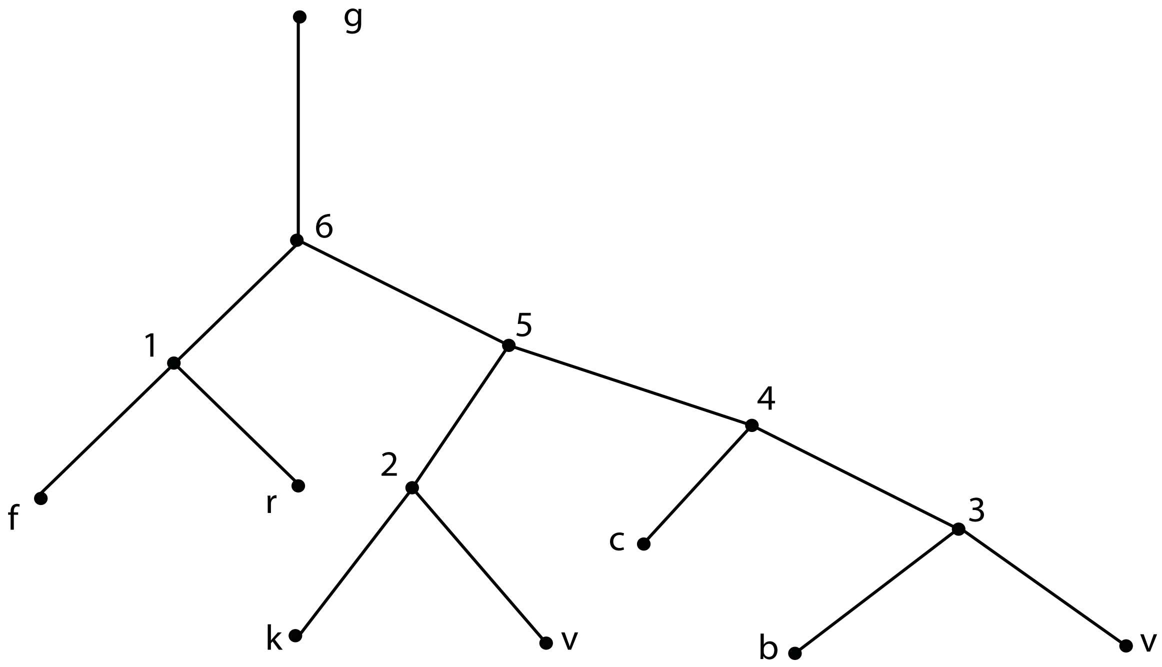


Figure 4.2: The tree used in the simulations. In the figure, nodes b, c, f, g, k, r, v and y represent the taxa of *P. berghei*, *P. chabaudi*, *P. falciparum*, *P. gallinaceum*, *P. knowlesi*, *P. reichenowi*, *P. vivax* and *P. yoelii*

log-likelihoods corresponding to local maxima. For each of the 94 set of estimates, we obtained the best-fitting NSGTR joint probability matrices. The best-fitting NSGTR joint probability matrices for the estimated permutation had much smaller distances between the estimates and true values than the best-fitting NSGTR matrices for other permutations. None of the 94 sets of best-fitting permutations corresponded to the original BH estimates reinforcing our results indicating that BH estimation alone is only accurate up to permutation (Zou et al. 2011a).

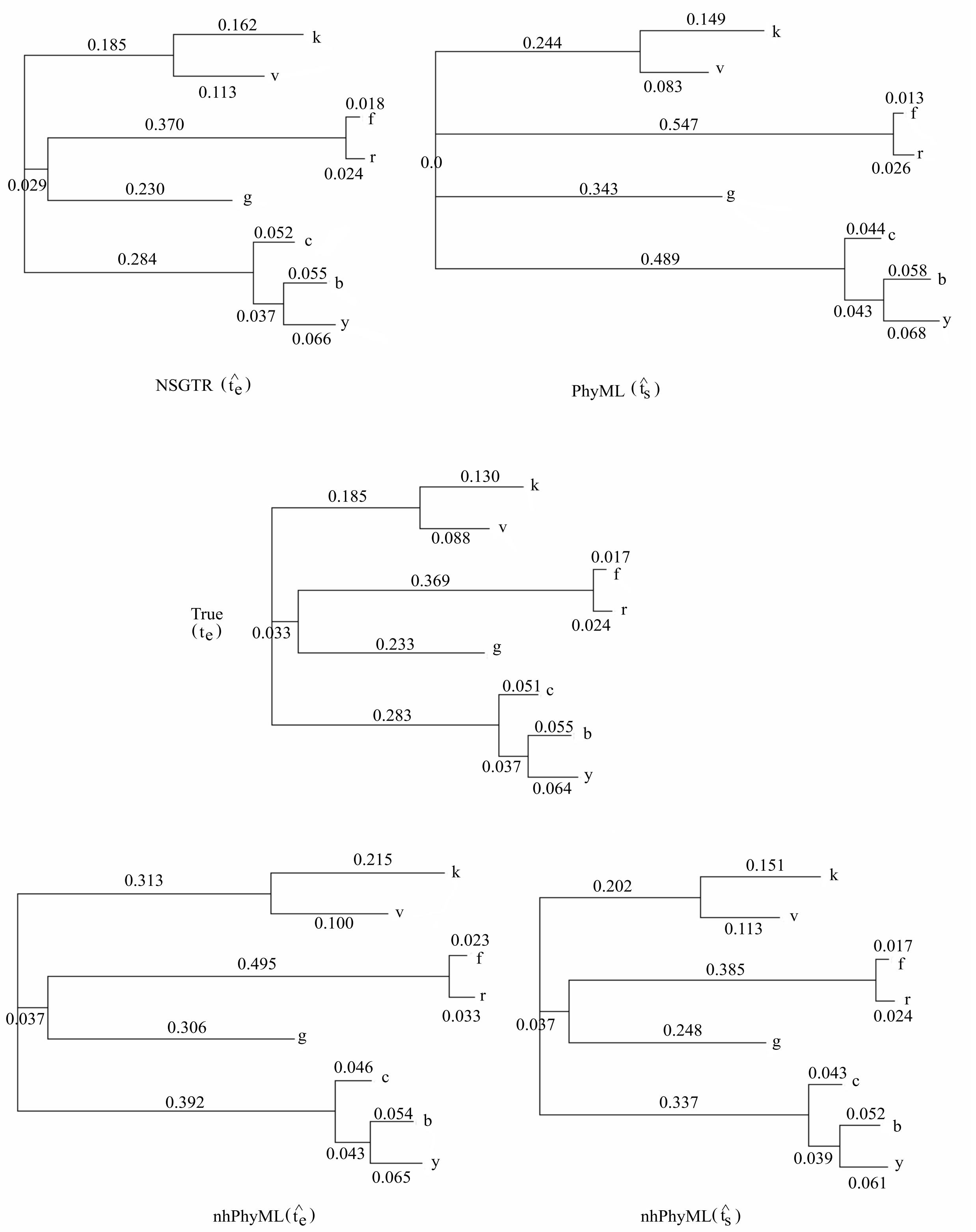
To test our algorithm for estimating the permutation, we considered the following: 1. whether there exists a distinct minimum sum of squares among 24 or 576 permutations for each edge and; 2. whether our algorithm can find the permutation which gives the distinct minimum sum of squares along edges. To address these questions, we estimated NSGTR models for all 24 permutations for the edges that connect an internal node and a leaf or the root node and 576 permutations for the edges that connect two internal nodes. Estimation was repeated in both evolutionary directions. Results for multiple datasets clearly showed that among all permutations of an edge, only one had a distinctly smaller sum of squares than the rest. The estimated permutations over the entire tree were those permutations which gave the smallest $minSS$ along edges. These tests were conducted for 18 simulated datasets under both mild and extreme parameter settings. In all cases, only one permutation gave a $minSS$ clearly smallest along all edges.

4.4.3 Edge Length Estimation

We compared the parameter estimates obtained from our NSGTR fit to the estimated BH model on the datasets with results estimated under a stationary GTR model using PhyML (Guindon and Gascuel 2003) and the GG nonstationary model implemented in nhPhyML (Boussau and Gouy 2006). Estimated edge lengths from PhyML and nhPhyML are the t_s parameter using the conventional rescaling for stationary models, $-\sum_j \Pi_j R_{jj} = 1$, where Π is the stationary frequency vector. For the GTR model, this t_s is equivalent to the expected number of substitutions per site. However, for the nonstationary evolutionary processes accommodated by the GG or the BH models along edges, the standard t_s edge length parameter is not the expected number of substitutions; the latter (i.e., t_e) is instead correctly computed using equation 4.4. In our experiments, we obtained estimates of the t_s parameter edge lengths from GTR and GG using PhyML and nhPhyML. We also computed the estimates of the expected number of substitutions (the t_es) along edges using the estimates of joint probabilities of GG and NSGTR models in equation 4.4. Since this is a simulated dataset, we also present the true average number of substitutions in Figure 4.3.

For the tree in Figure 4.3, we can separate the edges into two groups: a long edge group containing edges (5, 2), (6, 1), (6, g) and (5, 4) and a short edge group containing edges (1, f), (1, r), (4, c), (3, b) and (3,y). The estimates of edge lengths for the short edge group are very similar across different models and methods for calculating edge lengths. In the long edge group, the NSGTR estimates tend to be the closest to the true edge lengths. The estimates of t output by nhPhyML t_s and the expected number of substitutions for this dataset, t_e, obtained by correcting the t_s parameter with equation 4.4 did much better than the estimates of the GTR model from PhyML. For the GTR model, the estimates for all edges were poor, and especially for edge (6, 5) that was estimated to be zero when its true value was 0.033. Notably, GTR model estimates stretched all edges in the long edge group. That the GTR model performed worst under these conditions is not surprising given that it was badly misspecified.

We did an additional simulation to explore the effects of edge lengths estimated by nhPhyML under a correctly specified GG model. Using INDELible, we simulated a series of pairs under the GG model with 77313 sites. All pairs had the same starting state frequency vector $\{0.41, 0.10, 0.16, 0.33\}$ for character states A, C, G, T, G+C-content of 0.9 and transition/transversion ratio parameter 2. The edge lengths in



NSGTR ($\hat{t}_e$)

PhyML ($\hat{t}_s$)

True (t_e)

nhPhyML($\hat{t}_e$)

nhPhyML($\hat{t}_s$)

Figure 4.3: Estimates of edge lengths under four models of simulated dataset with extreme parameters

parameters are different for pairs. The results are presented in Table 4.1. As can be seen, the estimated expected number of substitutions t_e are much closer to the true values than the t_s edge lengths output by nhPhyML.

Table 4.1: The edge lengths and expected numbers of substitutions in the simulation for testing nhPhyML estimates of edge lengths

t_s		t_e	
True*	Estimate*	True**	Estimate**
0.1	0.159	0.115	0.141
0.2	0.293	0.224	0.260
0.3	0.410	0.327	0.361
0.4	0.508	0.426	0.451
0.5	0.624	0.520	0.569
0.6	0.720	0.610	0.630
0.7	0.806	0.697	0.798
0.8	0.909	0.782	0.793
0.9	0.996	0.864	0.880
1.3	1.377	1.176	1.199
1.4	1.472	1.251	1.203
2.0	2.080	1.683	1.333

True*: edge length in T92 model
Estimate*: the estimates of nhPhyML
True**: the true expected numbers of substitutions
Estimate**: the estimates of the expected numbers of substitutions

4.4.4 The BH Model is Useful when Evolutionary Processes are Non-stationary

The log likelihoods for the GTR and GG models were much smaller than for the BH model for a dataset simulated using the NSGTR estimates from the *Plasmodium* dataset. The log likelihoods for the GTR, GG, and BH models ranged from -412260 to -395567. As there are 8 taxa in the *Plasmodium* dataset, for a fixed topology, there are 22 free parameters using the GTR model, 30 free parameters in the GG model and 123 free parameters in the BH model. The differences of log likelihoods and degrees of freedom between the BH and GTR models are 16693 and 110 respectively and the corresponding differences between the BH and GG models are 5732 and 93 respectively; likelihood ratio tests with p-values close to zero clearly reject the simpler models in favour of BH. We also computed the sums of squares of differences between true values and estimates of expected numbers of substitutions, frequency vectors at nodes and transition matrices along edges. The sum of squared differences for the expected number of substitutions t_e were 0.005, 0.012 and 0.013 for the NSGTR, nhPyML and PhyML estimates; the sums of squares for the frequencies at nodes were 0.04, 0.14, and 0.11 for NSGTR, nhPhyML and PhyML estimates. Similar results were obtained for the sums of squares for transition matrices and joint probability matrices. Comparing the results of estimates of the extreme parameters dataset of three models, we have sums of squares of 0.00006, 0.05 and 0.3 for the expected numbers of substitutions and 0.0012, 1.04 and 1.14 for the nodes frequencies. In all cases, the NSGTR estimates are closest to the true values.

Similarly, the NSGTR estimates of frequency vectors at all nodes were always the best amongst the models compared (see Figure 4.4). The PhyML (GTR) and nhPhyML (GG) estimates varied and did not consistently over-estimate or under-estimate frequencies. Looking at the various estimated frequencies at each node, the NSGTR results clearly agree more closely to the estimates of the BH model under the estimated permutations than both the nhPhyML and PhyML results. The results for the simulated dataset with extreme parameters are shown in Figure 4.5 with similar results obtained, leading to similar conclusion.

For the real *Plasmodium* data, the largest NSGTR edge length estimate is for edge (5, 2). This largest value also corresponds to the largest frequency changes with the sum of squares of the differences 1.9×10^{-2}. For other edges, the sums of squares of the differences of frequencies at two end nodes at most 6.3×10^{-3}. The edge $(5, 2)$ separates the taxa k, v from the others. The NSGTR estimates do well at fitting the changes in frequencies; the nhPhyML estimates also had the largest changes along edge (5, 2) but its estimates are not as close due to the GG model restriction that frequencies of states C and G be the same. Since there is only one frequency vector for the entire topology, unsurprisingly, PhyML was unable to accurately estimate the frequencies for nodes k, v and 2.

The results for the simulated dataset in Figure 4.5 show a more complex example with extreme parameters models along edges. In the true model, node g, which is the root in the NSGTR model, has a high A+T-content with character state A having the highest frequency. The cluster of nodes 1, f and r have high G+C-contents with C having the highest frequency. The cluster of nodes 2, k and v have roughly equal A+T-contents and G+C-contents with G having the highest frequency. Finally, the cluster of nodes 4, c, 3, b and y have high A+T-contents with T having the highest frequency. The estimates of NSGTR and the estimates of the BH under the estimated permutations were within 10^{-4} of the each other.

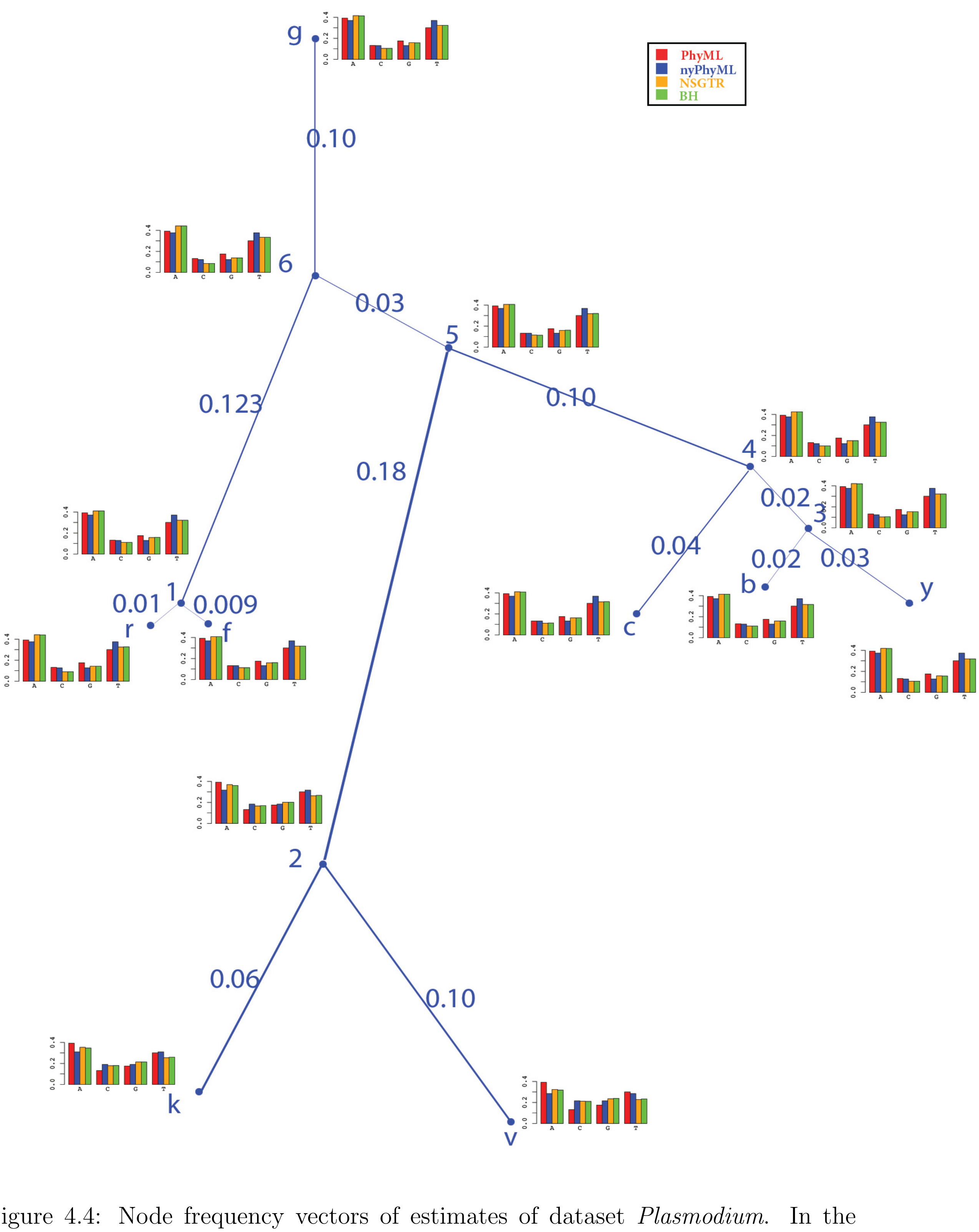


Figure 4.4: Node frequency vectors of estimates of dataset *Plasmodium*. In the figure, nodes b, c, f, g, k, r, v and y represent the taxa of *P. berghei*, *P. chabaudi*, *P. falciparum*, *P. gallinaceum*, *P. knowlesi*, *P. reichenowi*, *P. vivax*, and *P. yoelii*

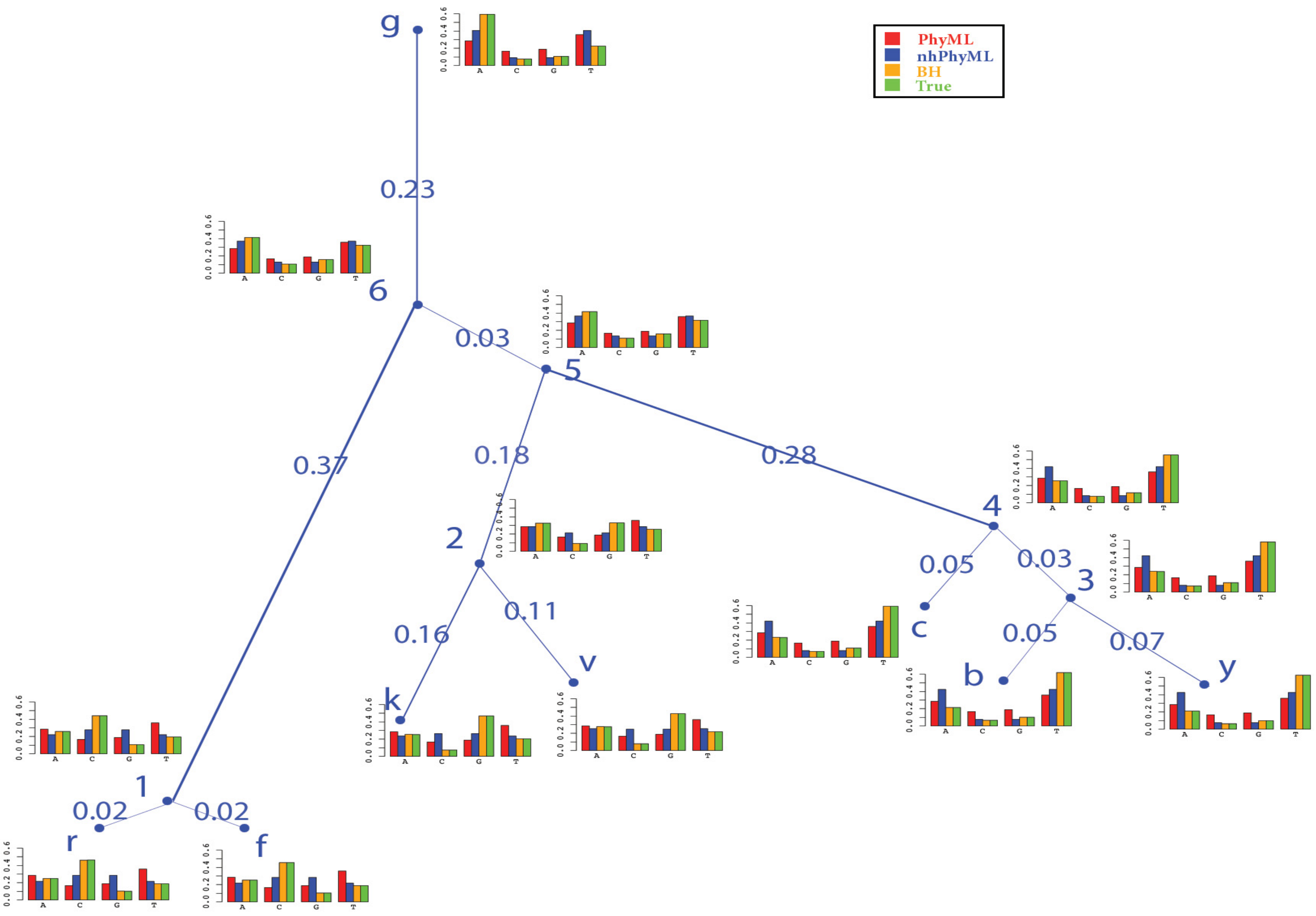


Figure 4.5: Node frequency vectors of estimates of simulated dataset using extreme parameters

Similar to the observed results for the *Plasmodium* data, large changes in frequency vectors are expected over longer evolutionary times. When the edge lengths are small, the frequency vector did not change much. For instance, the sum of squares of differences between the frequency vectors at nodes 1 and f is 0.0002 over an edge of length 0.02. In contrast, the sum of squared of differences between the frequency vectors at nodes g and f is 0.2636 over an edge of length 0.62. nhPhyML showed a similar pattern, with the frequency vector changes being largest along long edges although it did not provide accurate estimates of the frequencies. Again, estimates from PhyML could not accommodate the changing frequencies.

4.4.5 Discussion

When edge lengths are short, the estimates from GTR, GG, and NSGTR are quite similar. However, when a edge length was large, NSGTR estimates tended to estimate the changes of frequencies and edge lengths much better than the alternative methods. For the *Plasmodium* dataset shown in Figure 4.6, the edge lengths output from GTR and GG are not very different from expected numbers of the substitutions obtained using equation 4.4 and the estimates of NSGTR model. This is not surprising for this particular dataset because the NSGTR estimates of parameters for this dataset indicated that exchangeabilities and stationary frequencies did not change much over the tree except at the edge (5, 2). The processes along most edges are therefore close to being stationary processes. However, our simulation study under more extreme nonstationary parameter settings clearly shows the improved accuracy of the NSGTR method relative to the GG and GTR models.

While the NSGTR model estimated many parameters well, it did not do a good job at locating the root. Evaluating the $minSS$ values rooting at different nodes of *Plasmodium* dataset yielded two distinct values. Rooting at the nodes 1, 3, 4, 5, 6, b, c, f, g, r, and y gave a $minSS$ value of approximately $2.34E-04$ whereas rooting at the nodes 2, k, and v gave a $minSS$ value of approximately $1.48E-03$. This difference in $minSS$ allows us to rule out node 2, k, and v as the true root but does not distinguish between the others.

To explain our observations, we obtained the NSGTR substitution matrices for one of the rootings giving the minimum $minSS$. These NSGTR substitution matrices were used to obtain joint probability matrices which were given as input to the NSGTR routine. We used this routine to compute NSGTR models from true joint probabilities in the reverse directions along edges. We found that the true joint probability matrices coming from NSGTR in the reverse direction were almost identical (data not shown). Based upon these numerical results it appears that the evolutionary direction for a edge under the NSGTR model is not always recoverable. For the *Plasmodium* dataset, the estimates of edge lengths of the forwarding and reverse direction show few differences. However for the case of extreme datasets in our simulation, the direction effects were significant.

A natural follow-up question is whether there always exists a NSGTR model in the reverse direction if there exists a NSGTR model in the forward direction. We tested 18 datasets simulated under GTR models and BH models and obtained BH estimates. For each set of BH estimates, we examined 14 rooting positions. Among

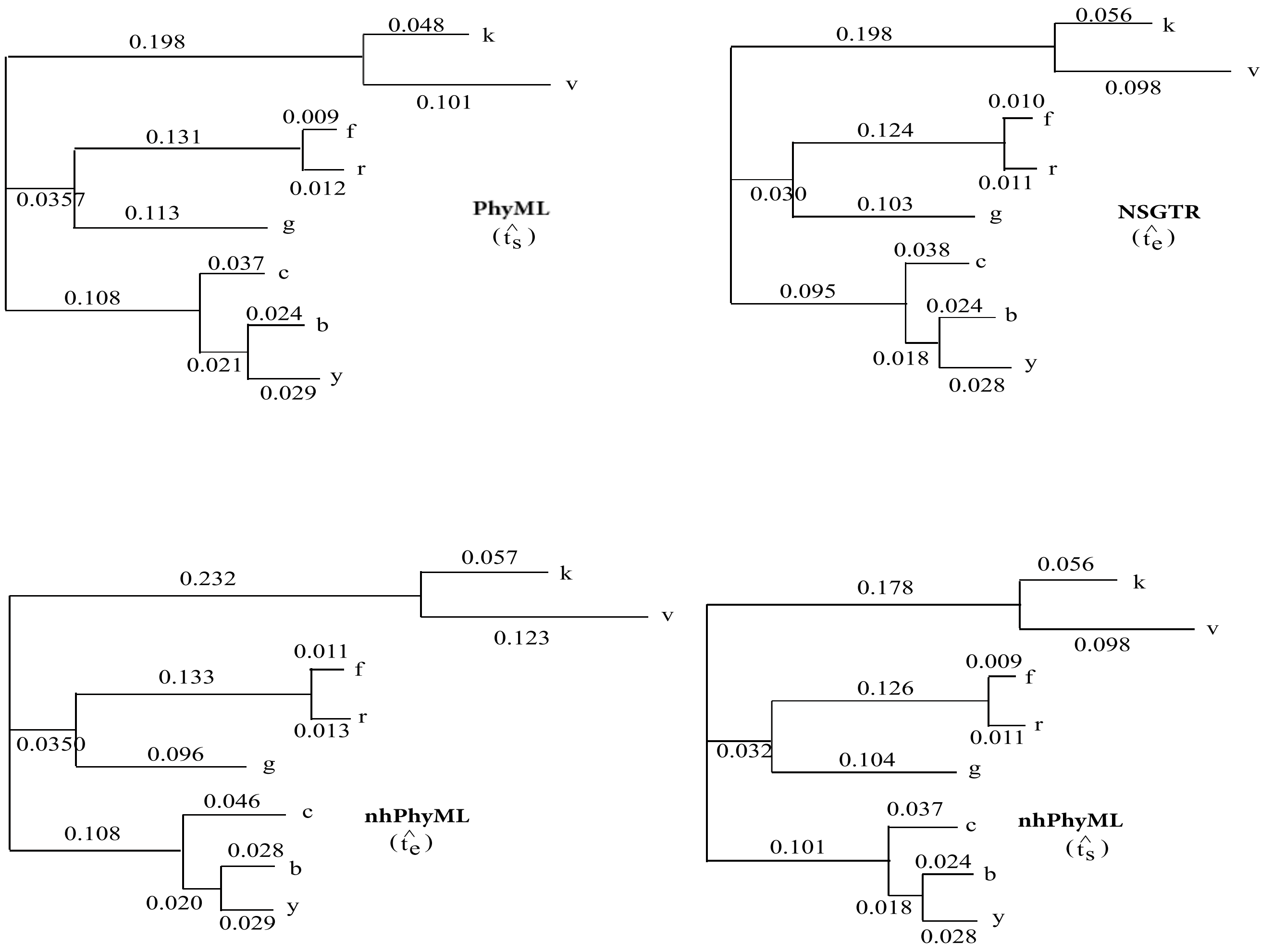


Figure 4.6: Estimates of edge lengths under four models of *Plasmodium* dataset. In the figure, nodes b, c, f, g, k, r, v and y represent the taxa of *P. berghei*, *P. chabaudi*, *P. falciparum*, *P. gallinaceum*, *P. knowlesi*, *P. reichenowi*, *P. vivax*, and *P. yoelii*

a total of 18*14=252 rooting trials, we only had two cases in which NSGTR models did not give exactly the same fit in both forward and backward directions.

4.5 Conclusions

The parameters of the BH model of Barry and Hartigan (1987) are joint probabilities along edges which have identifiability problems whereby multiple sets of estimates give the same likelihoods. Because of this, frequencies at internal nodes can not be correctly estimated although frequencies at leaf nodes will converge to true values as sequence length gets large. A further problem is that edge lengths are informative parameters but are not available from the BH model. By defining NSGTR models along edges, our algorithm finds the estimates of the transition matrices under the NSGTR model that best fits the BH estimates. Our simulations show that our algorithm is effective in resolving identifiability problems in both mild and extreme parameter settings.

In our solution, because the NSGTR model is nonstationary, non-standard methods were required to compute the edge lengths interpretable as expected numbers of substitutions along a edge. Our approach of using NSGTR estimates of the best-fitting BH estimates allows interpretable edge lengths to be estimated. The formulas given for edge length calculation are more broadly valuable for obtaining interpretable edge-lengths for all nonstationary models. For instance, the estimates of edge lengths currently given by the nhPhyML implementation of the GG model correspond to stationary model calculations but can be corrected using (4.4).

Chapter 5

The BH Mixture Model and its Applications

5.1 Introduction

The default general Markov model described by Barry and Hartigan (1987), known as the BH model, assumes that while evolutionary processes may vary across the tree, they remain constant across sites in DNA sequences. However, it has long been recognized that the evolutionary dynamics of sites differ depending on a variety of factors. For instance, in Fitch and Margoliash (1967), invariant sites in the amino acid sequences of *Cytochrome C* were identified. These sites thought to be fixed at the particular amino acid by purifying selection (i.e., any non-synonymous changes at these sites would cause the resulting protein to cease to function properly and the resulting mutant would have a strongly negative fitness effect). Such rates-across-sites variation also happens in the different codon positions where the third codon often has the fastest rate and the second codon has the slowest rate (Yang 1996). As pointed out by Semple and Taylor (2009), the physical structure of genome can also affect the rate of substitutions. Rates-across-sites variation is usually modeled by assigning rates to sites according to a discretized approximation to the Γ distribution (Yang 1994b) or a mixture of a discretized Γ distribution and an invariable site class, "I", (Gu et al. 1995) and has been shown to be important for accurate phylogenetic estimation (c.f. Yang 1994b, Thomas et al. 2006, Bromham 2009). In current applications of stationary models such as PAML(Yang 2007), PAUP* (Wilgenbusch 2003), PHYLIP (Felsenstein 1989), and TREE-PUZZLE(Schmidt et al. 2002), Γ models or Γ + I models have become widely used.

The YR nonstationary model in Yang and Roberts (1995) and the GG nonstationary model in Galtier and Gouy (1995) as implemented by Boussau and Gouy (2006) incorporated rate-across-sites variation using a discretized Γ distribution. In Jayaswal et al. (2007), a BH + I model was proposed. This model treated the invariable and variable sites differently through a mixture model. In a followup study, Jayaswal et al. (2011) suggested two stationary models in order to simplify the BH model. These two simplified models both allow invariable and variable sites.

While adjusting for rates-across-sites variation is now common practice in phylogenetic analyses, functional constraints on sites in a gene sequence can also change over time, causing shifts in site-specific evolutionary rates. This process is often referred to as heterotachy (Lopez et al. 2002) and has been modeled in various ways (c.f. Tuffley and Steel 1998, Galtier 2001, Huelsenbeck 2002, Susko et al. 2003). All of these models effectively allow rates and hence substitution matrices to vary across both sites and lineages (Wu and Susko 2010).

Here we introduce a BH mixture model that is more general than the BH+I model in that it not only allows completely different models along edges of a topology, but it also allows for different site classes whose evolutionary dynamics (e.g., rates, frequencies over sites and edges) can take any form. Such models will allow one to flexibly model situations, as described in Song (2010), for example, in which both rates and nucleotide frequencies vary across sites and/or lineages. Here, using a phylogenomic data set from *Plasmodium* species (Davalos and Perkins 2008), we show that the BH mixture model yields a much better fit to real data than the BH and BH + I models.

5.2 The Model

An example of the type of variation that the BH mixture model can capture is given in Figure 5.1. The illustration in Figure 5.1 is actually under the assumption that nonstationary GTR models, referred to as the NSGTR models, in Zou et al. (2011a) hold for different classes, which is a special case of the BH mixture model. A NSGTR model is the continuous Markov model which best fit into the transition matrix obtained from the estimate of the joint probability matrix of the BH model along the edge and the direction. Thus, for that edge and that direction, the edge length can be estimated. For the conventional GTR+Γ model as shown in Figure 5.2, rate multipliers for different classes are selected by using the means or medians of the quantile ranges of equally weighted categories under a Γ distribution. The trees in the different classes of a GTR+Γ model are constructed from the same tree by stretching or shrinking all edge lengths using a class-specific rate multiplier and sharing the same GTR model. By contrast, the BH mixture model allows completely different BH models in each class. It thus contains mixture-NSGTR and GTR + Γ + I models as special cases.

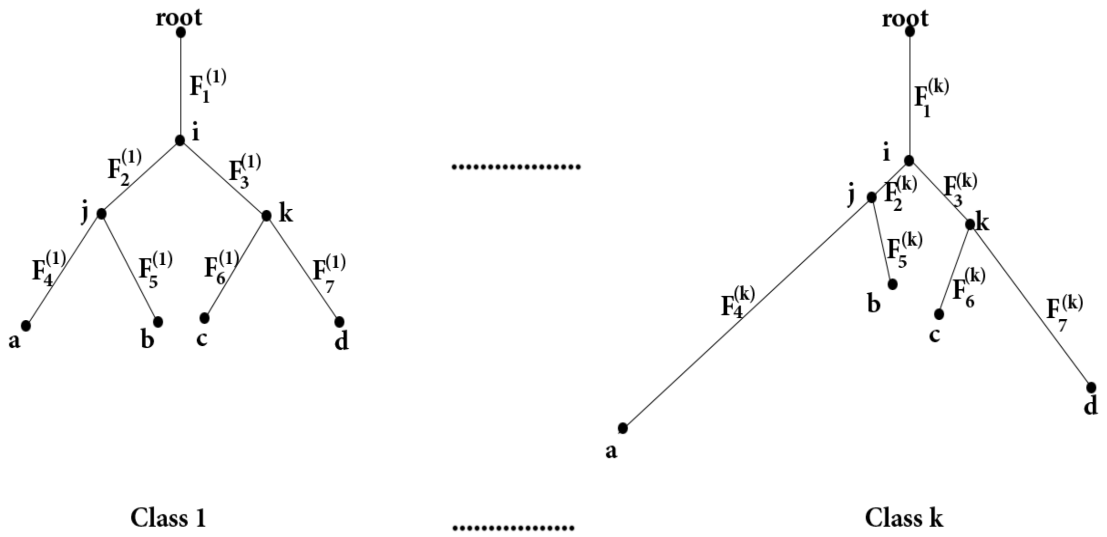


Figure 5.1: An illustration of the K-Class mixture model. Here each $F_e^{(k)}$ is a joint probability matrix for data at the end nodes of the edge. For illustration, edge lengths have been added as might arise if each $F_e^{(k)}$ corresponded to a differing GTR model, which we refer to as the NSGTR model.

The BH mixture model assumes that K classes of sites arise independently with probabilities $\underline{\omega} = \{\omega_1, \ldots, \omega_K\}$, where $\sum_{j=1}^{K} \omega_j = 1$. In each class, the evolutionary model is a BH model. For the j^{th} class and edge (a, b), let $F_{(a,b)}^{j}$ denote the joint probability matrix whose rs^{th} entry is the probability of observing nucleotides r and s at nodes a and b. For the j^{th} class, let F^j consist of the set of joint probability matrices for the $(2m - 3)$ edges of an m-taxon unrooted tree and write $F = \{F^1, \ldots, F^K\}$. Let the probability of observing site pattern $\underline{x}$ given class j be

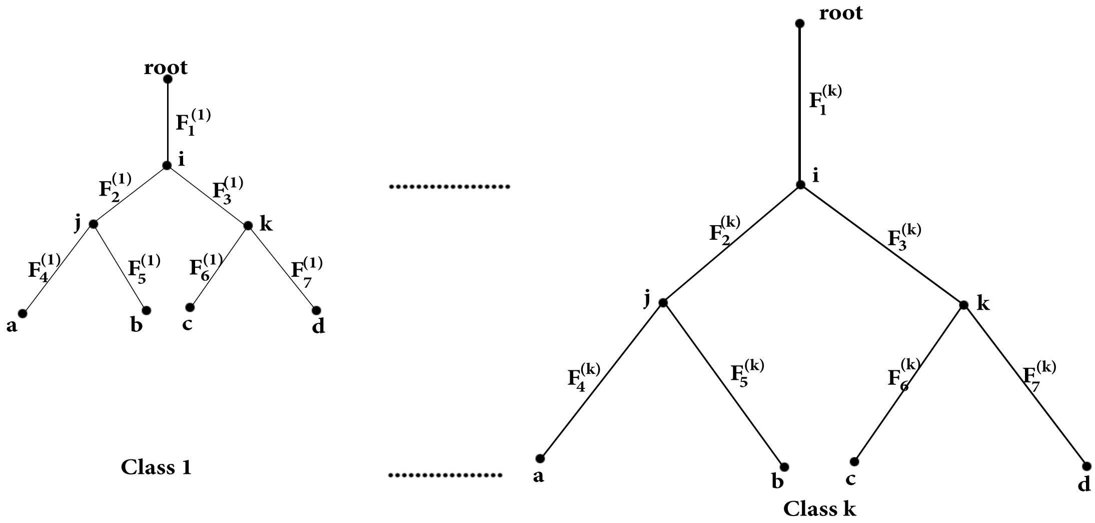


Figure 5.2: An illustration of the K-Class conventional Gamma+GTR model

denoted by $P(\underline{x}|F^j)$. The marginal probability of site pattern $\underline{x}$ is then obtained through summation

$$P(\underline{x}|F, \underline{\omega}) = \sum_{j=1}^{K} P(\underline{x}|F^j)\omega_j \tag{5.1}$$

Here a site pattern $\underline{x}$ is the set of nucleotides observed at leaves (i.e., the taxa) of the tree. For instance, $\underline{x} = AAAG$ is a site pattern for four taxa, in which the first three taxa had an A and the fourth a G.

For a given class j, the BH model of Barry and Hartigan (1987) and Jayaswal et al. (2005) holds where each edge is allowed to have its own joint probability matrix $F^j_{(a,b)}$. The only additional constraint is internal consistency which requires that internal node frequencies be the same regardless of the edge from which they are calculated.

5.2.1 Invariable Sites

It is often the case that data sets contain many invariant sites: that is, sites where the same nucleotide is observed for all taxa. Because of functional constraints the sites may in fact be invariable: i.e., unable to change. A class corresponding to invariable sites requires far fewer parameters than a BH model and thus including an invariable sites class reduces model complexity as shown in Table 5.1. For instance, with 3 classes and 7 taxa, there are 407 parameters in a full 3-class BH mixture whereas there are 275 parameters if one of the 3 classes is an invariable sites class. More formally, if we add an invariable sites class into the BH mixture model framework, a special case of the K classes BH mixture model, which we refer to as the BHK model, becomes a $K-1$ classes of BH models plus a invariable class, which we refer to as the BH$(K-1)$ + I model. For the BH$(K-1)$ + I model, define class 0 as the invariable sites class and the diagonal matrix Π^0 has the diagonal elements $\{\pi_A^0, \pi_C^0, \pi_G^0, \pi_T^0\}$ of frequencies of the states in the invariable sites class and denote $F = \{F^1, \ldots, F^{K-1}\}$ as the set of joint probability matrices of variable site classes. Then the BH$(K-1)$ + I model can be written as below.

$$P(\underline{x}|F,\underline{\omega}) = \begin{cases} P(\underline{x}|\Pi^0)\omega_0 \quad + \quad \sum_{j=1}^{K-1} P(\underline{x}|F^j)\omega_j, & \text{if } \underline{x} \text{ is an invariant site.} \\ \qquad\qquad\qquad\quad \sum_{j=1}^{K-1} P(\underline{x}|F^j)\omega_j, & \text{if } \underline{x} \text{ is not an invariant site.} \end{cases} \tag{5.2}$$

5.2.2 The Number of Parameters in the BH Mixture Model

Assume we have an m-taxon DNA data set. In an unrooted tree, there are $2m-3$ edges, $m-2$ internal nodes and m leaf nodes. For the BH model, there are 15

Table 5.1: The number of distinct site patterns and the differences of number of parameters between BHK and BH(K-1)

taxa	site patterns	BH	2-class		3-class		Dif.
			BH + I	BH2	BH2 + I	BH3	
3	64	39	43	79	83	119	36
4	256	63	67	127	130	194	70
5	1024	87	91	175	179	263	84
6	4096	111	115	223	227	335	108
7	16384	135	139	271	275	407	132
8	65536	159	163	319	323	479	156

"Dif": The number of Differences between BHK and BH(K-1) + I models.

parameters in each joint probability matrix and thus $15 \cdot (2m-3)$ parameters for the $2m-3$ edges. However due to the internal node consistency requirement, there are 6 constraints for each internal node. Combining everything together, the number of parameters of a m-taxon data set for a BH model is

$$15 \cdot (2m-3) - (m-2) \cdot 6 = 24m - 33 \tag{5.3}$$

For the BHK model , the total number of parameters in the BH mixture model is multiplied by K and $K-1$ parameters are added for the mixture weights.

$$24m \cdot K - 33 \cdot K + K - 1 = 24mK - 32K - 1 \tag{5.4}$$

In a BH$(K-1)$ + I model, there are $K-1$ BH models and one reduced BH model with three free parameters. Thus the total number of parameters is

$$(24 \cdot m - 33) \cdot (K - 1) + K - 1 + 3 = 24m \cdot (K - 1) - 32K + 35 \tag{5.5}$$

Comparing equations (5.4) and (5.5), the BH$(K-1)$ + I model has $24m - 36$ parameters less than BHK model. The difference of the number of parameters between BH$(K-1)$ + I and BHK model doesn't depend on K and only depends on the number of taxa m. Table 5.1 gives the number of parameters in the models BH, BHK and BH$(K-1)$ + I for $K = 2$ or 3 and 3 to 8 taxa.

5.3 Parameter Estimation for the BH Mixture Model

We used the expectation-maximization (EM) algorithm when estimating the parameters of a mixture model. The advantage of using the EM algorithm is that it simplifies implementation by allowing us to use the previously developed single-BH model in Zou et al. (2011b).

The EM algorithm was first described in Dempster et al. (1977) and McLachlan and Krishnan (1997) describe many of the developments since then. It gives an algorithm for maximum likelihood (ML) estimation from incomplete data. In our case, we can consider our data as incomplete data because we never observe which site is from which class. Thus, the sequence data and class labels for sites would be the complete data and the actual observed data, the sequence data alone is the incomplete data. In the following sections, we develop an EM algorithm for

ML estimation of the parameters, the joint probability matrices F^j and the weight distribution of classes $\underline{\omega}$, of a BH mixture model.

5.3.1 The EM Algorithm

Let $X = \{\underline{x}_1, \ldots, \underline{x}_N\}$ be the observed data for the N sites, where $\underline{x}_i$ is the site pattern for site i. Let S_i be the unobserved class label for site i and $S = \{S_1, \ldots, S_N\}$. The complete data is $\{X, S\}$ while the observed data is X. Since the sites are assumed independent, we can compute the site log-likelihood and then sum over all sites to get the log-likelihood for the complete data.

$$\mathbb{E}\left[\log P(\underline{x}_i, S_i | F, \underline{\omega}) | F^c, \underline{\omega}^c\right] = \sum_{j=1}^{K} \log\left[P(\underline{x}_i | S_i = j, F, \underline{\omega}) P(S_i = j | \underline{x}, F^c, \underline{\omega}^c)\right] \quad (5.6)$$

where F^c and $\underline{\omega}^c$ are the current parameter estimates. The EM algorithm iteratively updates F^{c+1}, $\underline{\omega}^{c+1}$ from F^c and $\underline{\omega}^c$ through

$$\{F^{c+1}, \underline{\omega}^{c+1}\} = \underset{\{F, \underline{\omega}\}}{\arg\max} \sum_{i=1}^{N} \mathbb{E}\left[\log P(\underline{x}_i, S_i | F, \underline{\omega}) | F^c, \underline{\omega}^c\right] \quad (5.7)$$

until the difference between $\{F^{c+1}, \underline{\omega}^{c+1}\}$ and $\{F^c, \underline{\omega}^c\}$ is small. The actual log-likelihood for the observed data of site i is $\log P(\underline{x}_i | F, \underline{\omega})$. Desirable properties of the EM-algorithm shown in Dempster et al. (1977) include that the log-likelihood at step $c + 1$ is greater than or equal the log-likelihood at step c and that it is guaranteed to converge to a local, but not necessarily global, maximum. In the next section, we will give details of how to compute $P(S_i | \underline{x}_i, F^c, \underline{\omega}^c)$ and how to update $\{F^{c+1}, \underline{\omega}^{c+1}\}$ iteratively.

5.3.2 Notation

5.3.2.1 E-step

The goal of the E-step is to calculate $\mathbb{E}\left[\log P(X,S|F,\underline{\omega})|F^c,\underline{\omega}^c\right]$. The log-likelihood of complete data can be expressed as

$$\log P(X,S|F,\underline{\omega}) \quad = \quad \sum_{i=1}^{N}\sum_{j=1}^{K} I\{S_i=j\}\log\left[P(\underline{x}_i|F^j)\omega_j\right]$$

Given F^c and $\underline{\omega}^c$, taking the expectation with respect to S gives

$$\begin{aligned}\mathbb{E}\left[\log P(X,S|F,\underline{\omega})|F^c,\underline{\omega}^c\right] &= \sum_{i=1}^{N}\sum_{j=1}^{K}\mathbb{E}\left[I\{S_i=j\}|\underline{x}_i,F^c,\underline{\omega}^c\right]\log\left[P(\underline{x}_i|F^j)\omega_j\right] \\ &= \sum_{i=1}^{N}\sum_{j=1}^{K}P(S_i=j|\underline{x}_i,F^c,\underline{\omega}^c)\log P(\underline{x}_i|F^j)\omega_j) \quad (5.8)\end{aligned}$$

Given $\underline{x}_i$, F^c and $\underline{\omega}^c$, we can compute

$$P(S_i=j|\underline{x}_i,F^c,\underline{\omega}^c) = \frac{P(\underline{x}_i|S_i=j,F^c,\underline{\omega}^c)P(S_i=j|F^c,\underline{\omega}^c)}{P(\underline{x}_i|F^c,\underline{\omega}^c)} = \frac{P(\underline{x}_i|F^{jc})\omega_j^c}{P(\underline{x}_i|F^c,\underline{\omega}^c)} \quad (5.9)$$

where $F^{jc}=\{F^{jc}_{(a,b)}|(a,b)\in\mathcal{E}\}$ is the current estimates of joint probability matrices of j^{th} class; $P(\underline{x}_i|F^{jc})\omega_j^c$ is the site likelihood of $\underline{x}_i$ in class j under current parameter F^c and $\underline{\omega}^c$ and $P(\underline{x}_i|F^c,\underline{\omega}^c)=\sum_{j=1}^{K}P(\underline{x}_i|F^{jc})\omega_j^c$.

5.3.2.2 M-step

In this step, parameters $\{F,\underline{\omega}\}$ maximizing $\mathbb{E}\left[\log P(X,S|F,\underline{\omega})|F^c,\underline{\omega}^c\right]$ are determined. We first consider maximization with respect to ω_j.

Updating ω_j

By the method of Lagrange multipliers, the maximizer of $\mathbb{E}\left[\log P(X,S|F,\underline{\omega})|F^c,\underline{\omega}^c\right]$ is also the maximizer of

$$\begin{aligned}
& \mathbb{E}\left[\log P(X,S|F,\underline{\omega})|F^c,\underline{\omega}^c\right] + \lambda(1-\sum_{j=1}^{K}\omega_j) \\
= & \sum_{i=1}^{N}\sum_{j=1}^{k} P(S_i=j|\underline{x}_i,F^c,\underline{\omega}^c)\log\left[P(\underline{x}_i|F^j)\omega_j\right] + \lambda(1-\sum_{j=1}^{K}\omega_j)
\end{aligned}$$

for some $\lambda > 0$ subject to the constraint that $\sum_{j=1}^{k}\omega_j = 1$. Taking the derivative with respect to ω_j and setting it to zero gives

$$\frac{1}{\omega_j}\sum_{i=1}^{N} P(S_i=j|\underline{x}_i,F^c,\underline{\omega}^c) - \lambda = 0$$

Thus ω_j is proportional to $\sum_{i=1}^{N} P(S_i=j|\underline{x}_i,F^c,\underline{\omega}^c)$. Since $\sum_j \omega_j = 1$, this gives

$$\omega_j = \frac{\sum_{i=1}^{N} P(S_i=j|\underline{x}_i,F^c,\underline{\omega}^c)}{\sum_{i=1}^{N}\sum_{j=1}^{k} P(S_i=j|\underline{x}_i,F^c,\underline{\omega}^c)} \tag{5.10}$$

Updating F^j

The expected complete log-likelihood $\mathbb{E}\left[\log P(X,S|F,\underline{\omega})|F^c,\underline{\omega}^c\right]$ can be written as

$$\sum_{i=1}^{N}\sum_{j=1}^{K} P(S_i=j|\underline{x}_i,F^c,\underline{\omega}^c)\log P(\underline{x}_i|F^j) + \sum_{i=1}^{N}\sum_{j=1}^{K} P(S_i=j|\underline{x}_i,F^c,\underline{\omega}^c)\log\omega_j$$

Since the only term involving F^j is $\sum_{i=1}^{N} P(S_i = j|\underline{x}_i, F^c, \underline{\omega}^c) \log P(\underline{x}_i|F^j)$, then

$$F^j = \arg\max_{F^j} \sum_{i=1}^{N} P(S_i = j|\underline{x}_i, F^c, \underline{\omega}^c) \log P(\underline{x}_i|F^j) \tag{5.11}$$

Equation (5.11) implies that updating F^j in an iteration of the EM algorithm is the same as ML estimation in the usual BH model except that $\log P(\underline{x}_i|F^j)$ is weighted by the non-integer $P(S_i = j|\underline{x}_i, F^c, \underline{\omega}^c)$.

Updating Π^0

BH$(K-1)$ + I is a special case of BHK where one class corresponds to an invariable sites class. Recall that when considering the invariable site class, the diagonal matrix Π^0 are the diagonal elements of the frequencies of invariable sites $\{\pi_A^0, \pi_C^0, \pi_G^0, \pi_T^0\}$; $F^c = \{F_1^c, \ldots, F_{K-1}^c\}$ is the set of joint probability matrices of $(K-1)$ variable site classes. The expected complete log-likelihood can be written as

$$\sum_{i=1}^{N} P(S_i = 0|\underline{x}_i, \Pi^{0c}, F^c, \underline{\omega}^c) \log\left[P(\underline{x}_i|\Pi^0)\omega_0\right] + \sum_{i=1}^{N}\sum_{j=1}^{K} P(S_i = j|\underline{x}_i, F^c, \underline{\omega}^c) \log\left[P(\underline{x}_i|F^c, \underline{\omega}^c)\omega_j\right] \tag{5.12}$$

Since only the first term in (5.12) involves Π^0, using the method of Lagrange multiplier, the maximizer of $\mathbb{E}\left[\log P(X, S|F, \underline{\omega})|F^c, \underline{\omega}^c\right]$ is also the maximizer of

$$\sum_{i=1}^{N} P(S_i = 0|\underline{x}_i, \Pi^{0c}, F^c, \underline{\omega}^c) \log\left[P(x_i|\Pi^0)\omega_0\right] + \lambda(1 - \sum_{l=1}^{4} \pi_l^0)$$

for some $\lambda > 0$ and subject to the constraint $\sum_{k=0}^{4} \pi_k^0 = 1$. For nucleotide data there four possible invariable site patterns, $\underline{x}_i = \text{AA}\ldots\text{AA}, \ldots, \underline{x}_i = \text{TT}\ldots\text{TT}$. Let $\{y_1, ..., y_4\}$ denote these patterns. Then Π^0 is the maximizer of

$$\sum_{l=1}^{4} n_l P(S_1 = 0|\underline{x}_1 = y_l, \Pi^{0c}, F^c, \omega^c) \log\left[\pi_l^0 \omega_0\right] + \lambda(1 - \sum_{l=1}^{4} \pi_l^0)$$

where n_l is the number of times the site pattern y_l occurred among the x_i. Taking the derivative with respect to π_l^0, setting it to zero and solving gives

$$\pi_l^0 = \frac{n_l P(S_1 = 0|\underline{x}_1 = y_l, \Pi^{0c}, F^c, \underline{\omega}^c)}{\sum_{l=1}^{4} n_l P(S_1 = 0|\underline{x}_1 = y_l, \Pi^{0c}, F^c, \underline{\omega}^c)} \tag{5.13}$$

5.4 Leaf Node Frequencies

It is shown in Jayaswal et al. (2005) that the ML estimate of the BH model gives marginal probabilities of leaf node data that exactly match the observed proportions of times the nucleotides arose at that leaf. For the mixture model, we can also obtain that the leaf frequency vectors fit consistently with empirical data frequencies at leaf nodes. Let N be the sequence length and n_r^a be the frequency of the character r at node a. The ML estimate of the marginal probability π_r^a of character r at node a satisfies

$$\pi_r^a = \frac{n_r^a}{N} \tag{5.14}$$

A proof of this property is given in Appendix A5.1.

5.5 Permutation and Edge Length

The classes in a mixture model are the individual BH models. As shown in Zou et al. (2011a), the ML estimates of joint probability matrices of the BH model for a tree are not unique. Given a set of ML estimates, there exist permutations of the rows or columns of the joint probability matrices that will give exactly the same likelihood. This problem of identifiability holds for the BH mixture model as well. To estimate internal node frequencies, we applied the algorithm of Zou et al. (2011b) to estimate the permutation separately for each class in a mixture model. In brief, each estimate of joint probability matrices of the BH model has at least 24 permutations of its rows and columns that give the same likelihood. The algorithm of Zou et al. (2011b) searches for the set of permutations and nonstationary GTR (NSGTR) models that give the smallest sum of squared differences between joint probability matrix entries for the NSGTR model and the corresponding entries for a fitted BH model. Here it is applied separately to each of the classes to generate fitted NSGTR models for that class. The pattern probabilities for invariable sites are determined entirely by the frequencies of nucleotides. Since, for invariable sites, leaf node frequencies are the same as internal node frequencies and since leaf node frequencies are identifiable, the pattern probabilities for invariable sites do not suffer the identifiability problems that the BH model has.

While edge-lengths are not parameters of the BH mixture model, estimates of them can be obtained from the best-fit NSGTR algorithm for a class using the expected substitutions per site correction formula in Zou et al. (2011b). To present overall edge lengths, edge lengths are averaged over classes as follows:

$$\sum_{j \in S} \omega_j t_e^j \tag{5.15}$$

where t_e^j is the edge-length the e^{th} edge and j^{th} class.

5.6 Simulations

Simulations were conducted generating 10,000, 50,000 or 200,000 sites. In the discussion that follows, when not specified, the true mixture model employed is BH2 + I; the invariable sites arise with probability 0.1; two other classes are equally weighted and arise with probability 0.45; the sequence length is 200,000 sites. For the two BH classes, we chose one class to be C+G rich and the other to be A+T rich. BH class simulations were conducted employing NSGTR models. The stationary frequencies of the NSGTR models along edges, edge lengths and node frequencies are shown in Figures 5.3, 5.4, 5.5 and 5.6. For ML estimation, we used multiple randomly generated joint probability matrices along edges as starting values. For estimation, our software which builds on Jayaswal et al. (2005) is used for the mixture model, PhyML in Guindon and Gascuel (2003) is used for the stationary model and nhPhyML in Boussau and Gouy (2006) is used for estimating the *GG* model.

5.7 Results

5.7.1 Inferring GTR + Γ + I Models

As a first step, we generate data from a model close to a GTR+Γ2+I model for 4-taxon data sets. In particular, we have two BH models with the same stationary frequency vector {0.03, 0.425, 0.475, 0.07} and exchangeabilities {1.0, 2.0, 2.5, 3.0, 2.0, 2.5} for $a \leftrightarrow c$, $a \leftrightarrow g$, $a \leftrightarrow t$, $c \leftrightarrow g$, $c \leftrightarrow t$, $g \leftrightarrow t$ for the GTR models along edges, but with the edge lengths in one of the classes were set to 0.05 between two adjacent nodes whereas the edge lengths in the other class were 0.75 between leaf node and internal node and 0.05 between internal nodes. Thus, in this setting the two classes are made distinct solely by the edge lengths over which the data evolved. For this set of experiments, we simulated data sets with different invariable class weights ranging from 0.1 - 0.6 and the equal weight for the two variable classes. For estimation, we employed the BH2+I mixture model and the GTR+Γ2+I model. The BH2+I has 130 parameters (see Table 5.1) as compared to 13 parameters for GTR+Γ2+I. For data sets with different proportions of invariable sites, Table 5.2 gives results of likelihoods ratio tests of BH2+I against GTR+Γ2+I , each of which have 117 degrees of freedom. Since the null hypothesis is true here, the small p-values are a bit surprising. Inspection of the parameter estimates revealed that both BH2+I and GTR+Γ2+I models overestimated the weight of invariable sites when the true weights are small. However the GTR+Γ2+I model was more accurate than the mixture model in general in estimating the frequencies of invariable sites and edge lengths. Those results indicated that for data close to stationary model such as GTR+Γ2+I, the BH mixture may be overparameterized and may not provide better estimates.

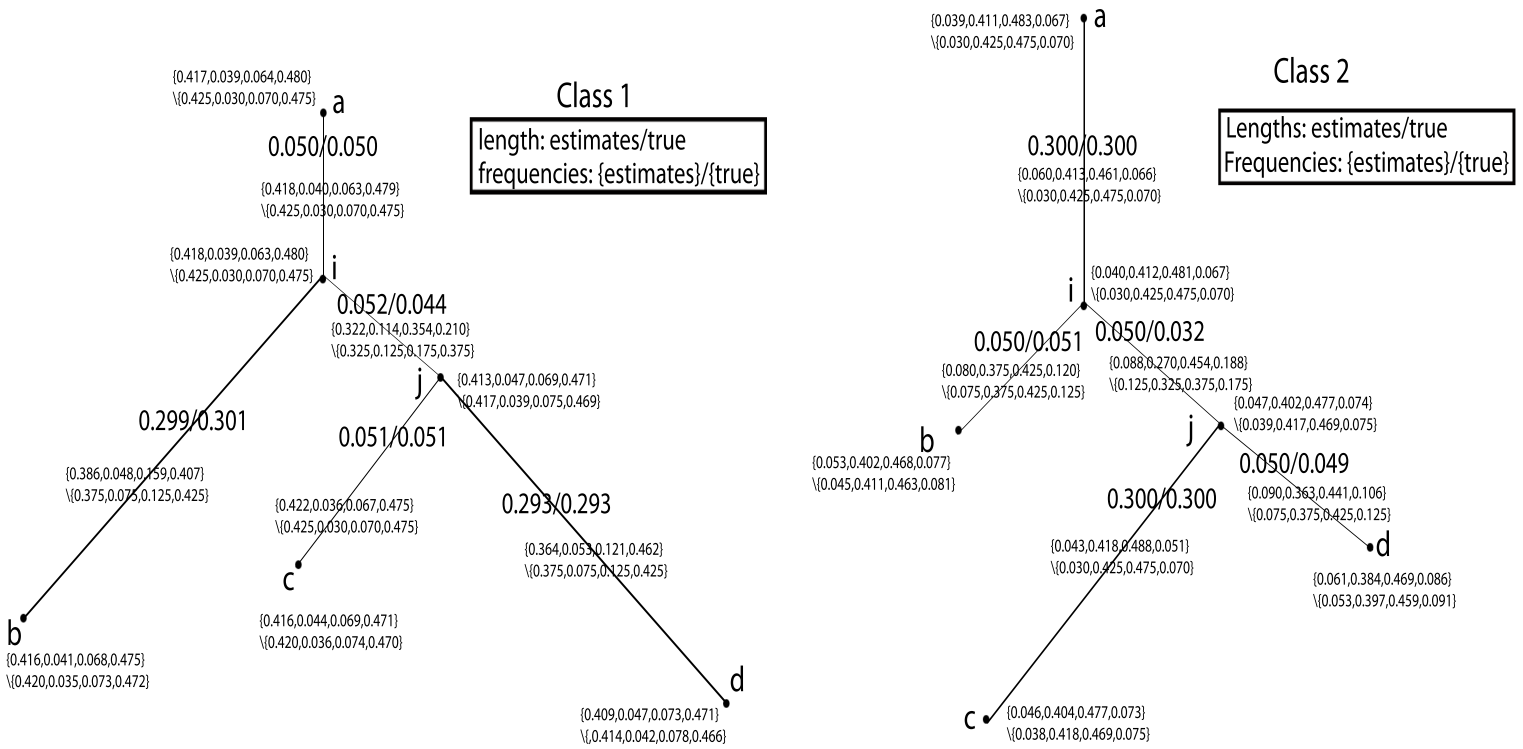
Table 5.2: The p-values of likelihood ratio tests for simulated data sets which had the models close to GTR + Γ2 + I.

proportion of invariable sites	0.1	0.2	0.3	0.4	0.5	0.6
p-values	0.004	0.0005	0.27	0.016	0.19	0.87

5.7.2 Identifying the Best Fitting Model

A data set was simulated under a BH2+I model using parameters of the four-taxon tree shown in Figure 5.3. In order to determine the best fitting model, this data set was fitted under GTR+ Γ2+I, GG+Γ2, BH, BH+I, BH2+I, and BH3 models.

For parameters set at their values used in the simulation, the log-likelihood for the BH2+I model for the 200,000 sites data set is -653008.48. The upper-bound on the log-likelihood is given by the log-likelihood of the empirical multinomial distribution model (i.e., a model where the probability of each site pattern is set as its empirical frequency) which in this case is -652459.7. To estimate the likelihoods for all of the other models, as mentioned before, we used multiple initial random starting values. The results presented are those which gave the best log-likelihoods under each of the BHK and BH$(K-1)$ + I models over all possible initial starting points. The overall results are shown in Table 5.3.



Tree 1: G + C rich

Tree 2: A + T rich

Figure 5.3: True values and estimates of parameters under BH2 + I Model of simulated four-taxon data set

Table 5.3: Results of likelihoods, class weights, and invariable frequencies obtained using estimates which had the largest likelihoods among different initial values for each model for the tree in Figure 5.3

Model	m. lk	likelihoods	w.inv	w.c1	w.c2	w.c3	inv.a	inv.c	inv.g	inv.t
True Values	-652459.72	-653008.48	0.100	0.450	0.450	na	0.200	0.300	0.100	0.400
GTR+Γ2+I	-652459.72	-670488.75	0.030	0.485	0.485	na	0.228	0.230	0.252	0.290
GG+Γ2	-652459.72	-686654.62	na	0.500	0.500	na	na	na	na	na
BH + I	-652459.72	-654353.25	0.066	0.934	na	na	0.164	0.369	0.033	0.434
BH2	-652459.72	-652689.73	na	0.462	0.538	na	na	na	na	na
BH2 + I	-652459.72	-652535.98	0.106	0.445	0.449	na	0.182	0.319	0.096	0.403
BH3	-652459.72	-652540.43	na	0.401	0.301	0.298	na	na	na	na

m.lk: the loglikelihood calculated from pattern frequencies of simulated dataset.
w.inv/c1/c2/c3: the class weights of invariable sites, class one, two and three.
inv.a/c/g/t: the frequencies of invariable sites which has all character “A”s, “C”s, “G”, and “T”s.

Using likelihood ratio tests, we find that the improvements of BH2+I model against GTR+Γ2+I and GG+Γ2 are significant. Further tests showed that the improvements of BH+I against BH, BH2 against BH+I, BH2+I against BH2 are all significant. In contrast, the likelihood ratio test comparing the BH2+I and BH3 fitted models was not significant. From our tests, we can conclude that the under-parameterized models such as GTR+Γ2+I, GG+Γ2, BH, BH+I, and BH2 would not be chosen in favor of the true model and by contrast, while the over-parameterized BH3 model gave greater log-likelihoods than the correct model, this improvement was not significant. Thus this simulation confirmed that we will be able to correctly determine the true model with a large number of sites.

Beyond likelihood ratio tests, we can also see that the estimates of weights of classes, frequencies of invariable sites, node frequencies, edge lengths, stationary frequencies in the NSGTR models along edges of the best fitting BH2+I model are reasonably accurate, and amongst all models examined are the closest ones to the true values. For instance, in the simulation, the probability of invariable sites was set to 0.1 and the probability for each of classes 1 and 2 was 0.45. The corresponding estimates of weights under the BH2+I model was 0.106 for invariable sites, 0.445 and 0.449 for classes 1 and 2 respectively. The estimates of edge lengths shown in Table 5.4 are the averaged edge lengths considered over all classes using (5.15) compared with the outputs of PhyML and nhPhyML, both of which are the expected substitution per sites-based edge lengths under conventional stationary model. The edge length estimates under the stationary GTR+Γ2+I model and the

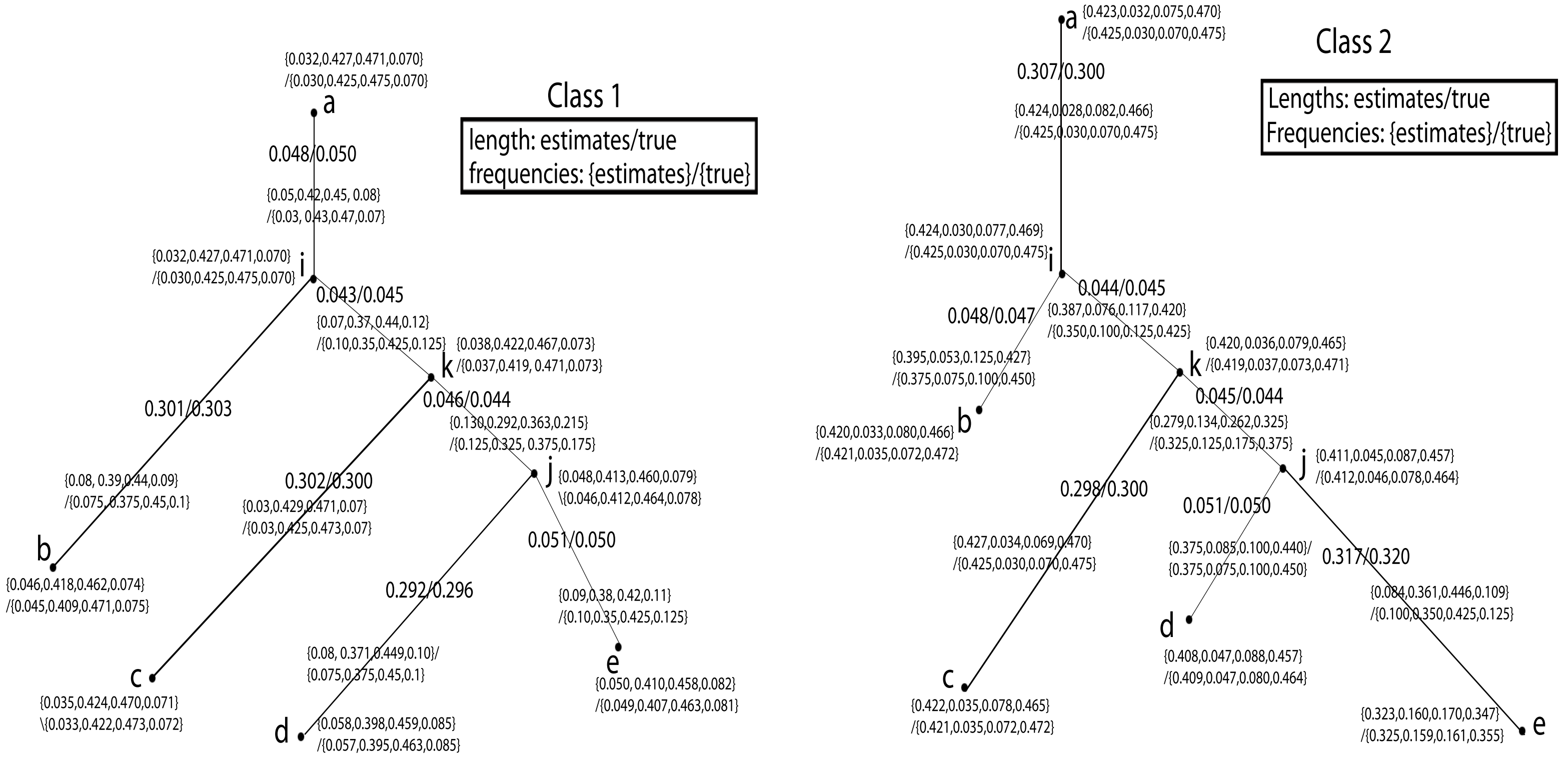
GG+Γ2 model are very similar and both are the least accurate estimates of all models examined. In this particular case, except for the estimates generated by the BH2+I model, almost all estimates underestimated the edge lengths. However there is no evidence showing that the edge lengths under wrong model have to be underestimated or overestimated. Estimates of node frequencies in the different classes showed a similar pattern and again those provided by the 'correct' BH2+I model were the best.

We also compared the parameters for the best-fitting NSGTR models along edges with the original BH2+I fitted parameters. Considering the estimated stationary frequency vectors of NSGTR models along edges and node frequency vectors in individual classes, we found that the BH2+I node frequency vectors are more accurate than those estimated by the NSGTR models. This is not surprising because the NSGTR model fitting requires more calculations such as symmetrization and eigen decomposition and both of those steps may introduce extra errors into the estimates.

We have conducted similar simulations for five and six-taxon data sets. The NSGTR models along edges are shown in Figures 5.4 and 5.5. For these two sets simulated under the BH2+I model, we estimated them using GTR+$\Gamma + I$, GG, BH, BH+I, BH2, BH2+I, and BH3. Using the likelihood ratio test similar to four-taxon estimates, the true model, the BH2+I model, was identified for both five and six-taxon estimates. Also the estimates of the BH2+I model provided the most accurate estimation for the edge lengths, frequency vectors at nodes, the weights distribution,

Table 5.4: Estimates of weighted edge lengths under different models for the tree in Figure 5.3

Edge	TRUE	BH2 + I	GTR + Γ2 +I	GG + Γ2	BH	BH+I	BH2	BH3
ai	0.158	0.156	0.139	0.135	0.146	0.150	0.147	0.141
ib	0.157	0.156	0.146	0.139	0.146	0.149	0.159	0.149
jc	0.158	0.157	0.138	0.133	0.148	0.152	0.146	0.147
jd	0.154	0.153	0.147	0.139	0.141	0.144	0.159	0.145
ij	0.040	0.039	0.031	0.030	0.035	0.034	0.040	0.036



Tree 1: G + C rich

Tree 2: A + T rich

Figure 5.4: True values and estimates of parameters under BH2 + I model of simulated five-taxon data set

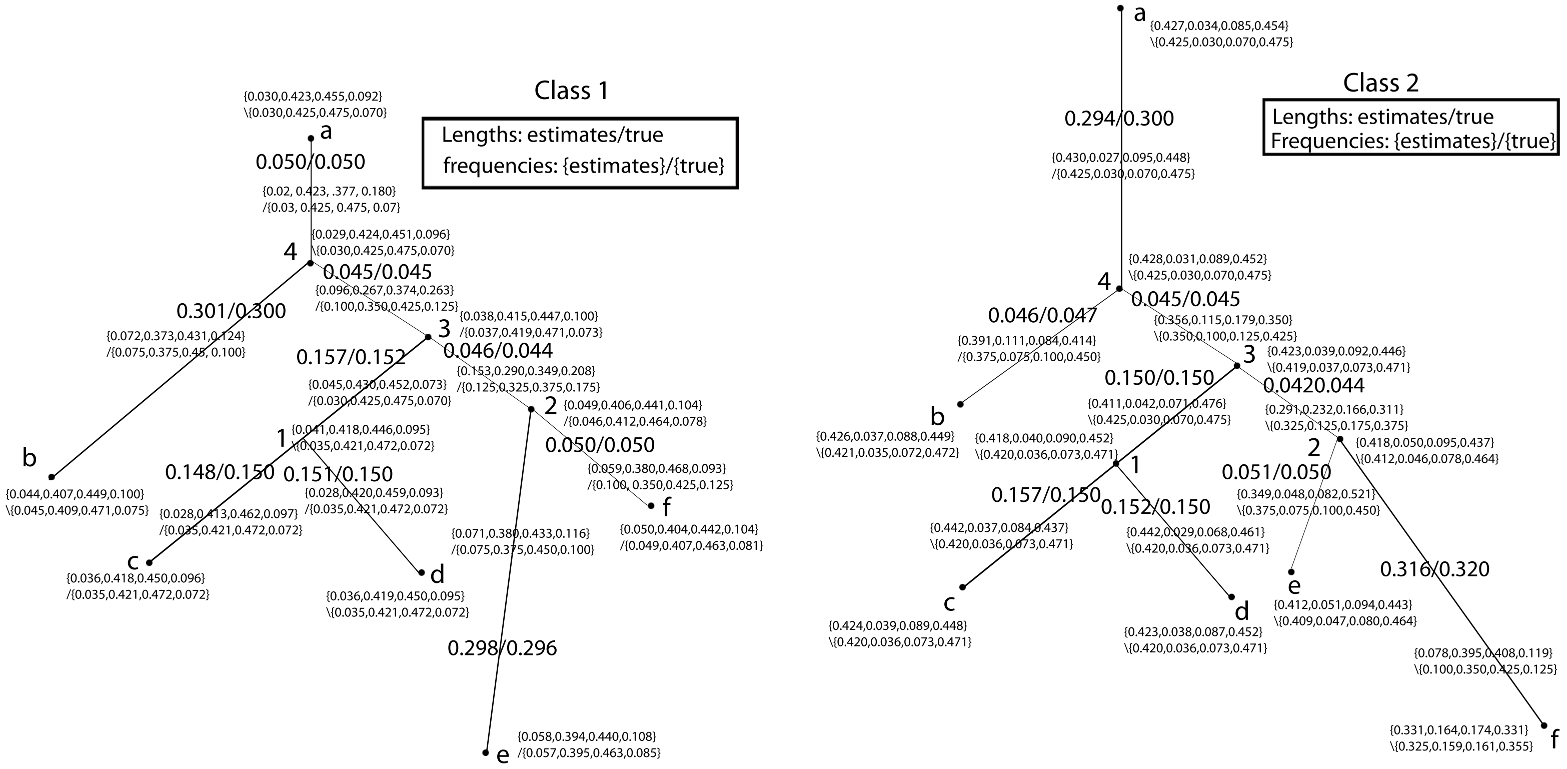
the stationary frequencies of NSGTR models along edges.

5.7.3 Estimation for Data Sets with More Taxa

The number of parameters in the mixture models are far greater than required for the BH model as shown in Table 5.1. As a consequence, our current implementation is still slow when optimizing and finding ML estimates with large data sets. Thus when extending our observations to larger data sets, we only tested cases with five, six, and seven taxa. All estimates were obtained under the true model, BH2+I, in order to find out how the larger number of taxa affects the estimates.

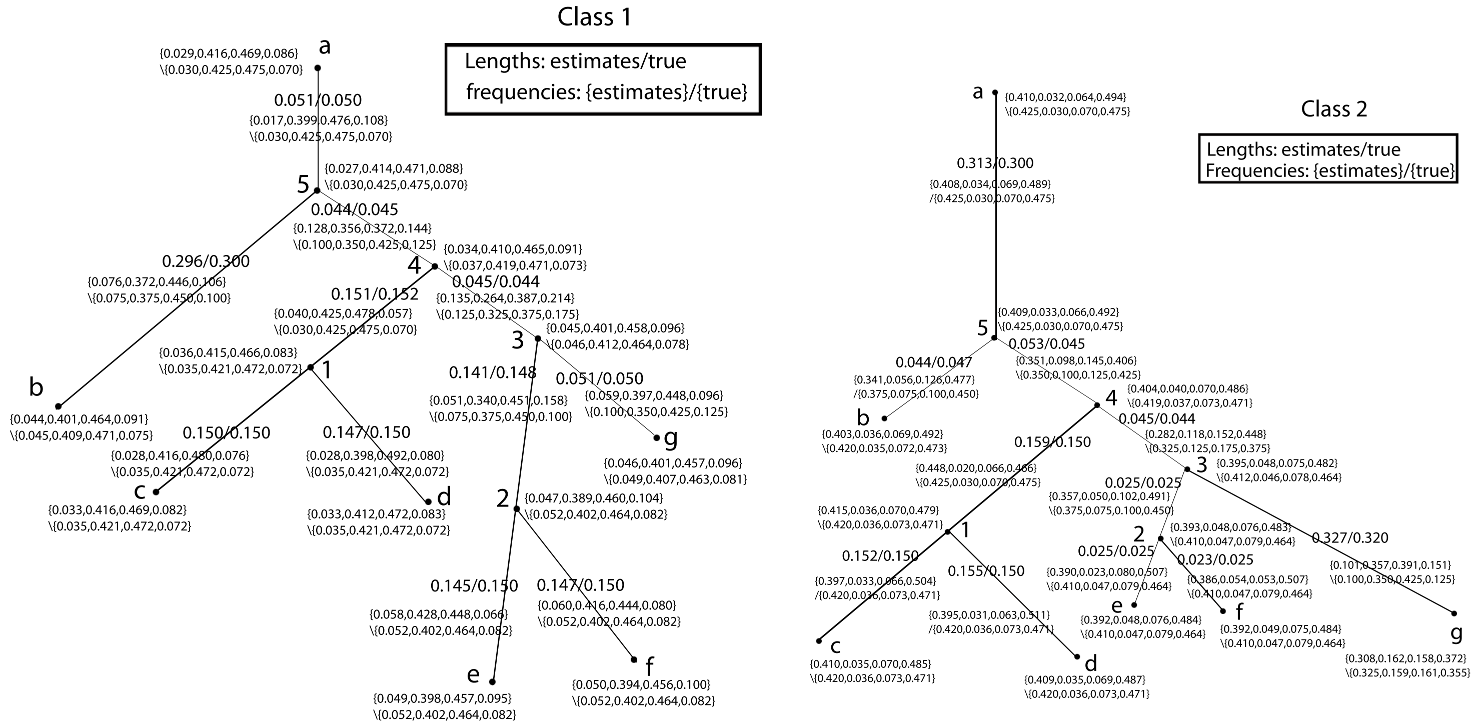
From our simulations, we found that the weighted node frequencies and expected number of substitutions were much closer to the true values than the class-specific parameters. When analyzing results, we separated our results into two groups. In the first group, we focus on the estimates from individual classes. As shown in Figures 5.3 - 5.6, the parameters of the NSGTR models along edges are different in individual classes. Those differences of parameters in individual classes are mimicking the different evolutionary processes of the various site types. Therefore, how those parameters are estimated is a focus of our analysis. Figures 5.3 - 5.6 show the results of estimates for individual classes. In the second group, we will look at the estimates that combined and weighted the estimates of different classes such as node frequencies and edge lengths. Figure 5.7 shows the results. The means of squares of the differences of the estimates of weighted node frequencies and the true values were at 10^{-5} or less; the means of squared biases of weighted frequencies were even less than this. A similar analysis for the true values and estimates of expected number of substitutions along edges showed similar results. When comparing the parameters class by class, the mean squared differences and the mean of biases are larger than weighted estimates. This observation indicated that as long as the weighted frequencies of leaf nodes converged to the empirical leaf nodes frequencies, the estimates of joint probabilities only need to satisfy the constraint such as the internal nodes consistency in addition to summing over all entries being equal to one in a joint probability matrix.

It is reasonable to expect that estimation will improve with more taxa (Zwickl and Hillis 2002). We explored this with an experiment in which a six-taxon tree was created by breaking the long edge $(2, c)$ of a five-taxon tree as shown in Figure 5.8. In the six-taxon tree, the parameters of the NSGTR models of the edges $(2, 1)$, $(4, c)$ and $(4, d)$ are exactly the same as the parameters of the NSGTR model of the edge $(2, c)$ in five-taxon tree; the node frequency vectors at nodes c and d are the same as the frequency vector of node c in five-taxon tree. It is well known that the long edges are hard to estimate accurately because, as in our case, the internal node 2 is farther from the observed data. By breaking up long edge $(2, c)$, we are adding more information when estimating node 2 as well other internal nodes. In this experiment, the sequence length is 50,000 sites and the true model is BH2+I. The results in Figure 5.9b show the boxplots of the differences between the estimates and true values of node frequencies of two classes individually. The boxplot labeled 5.c1 gives differences between five-taxon estimates and the corresponding true generating



Tree 1: G + C rich **Tree 2: A + T rich**

Figure 5.5: True values and estimates of parameters under BH2 + I model of simulated six-taxon data set



Tree 1: G + C rich

Tree 2: A + T rich

Figure 5.6: True values and estimates of parameters under BH2 + I model of simulated seven-taxon data set

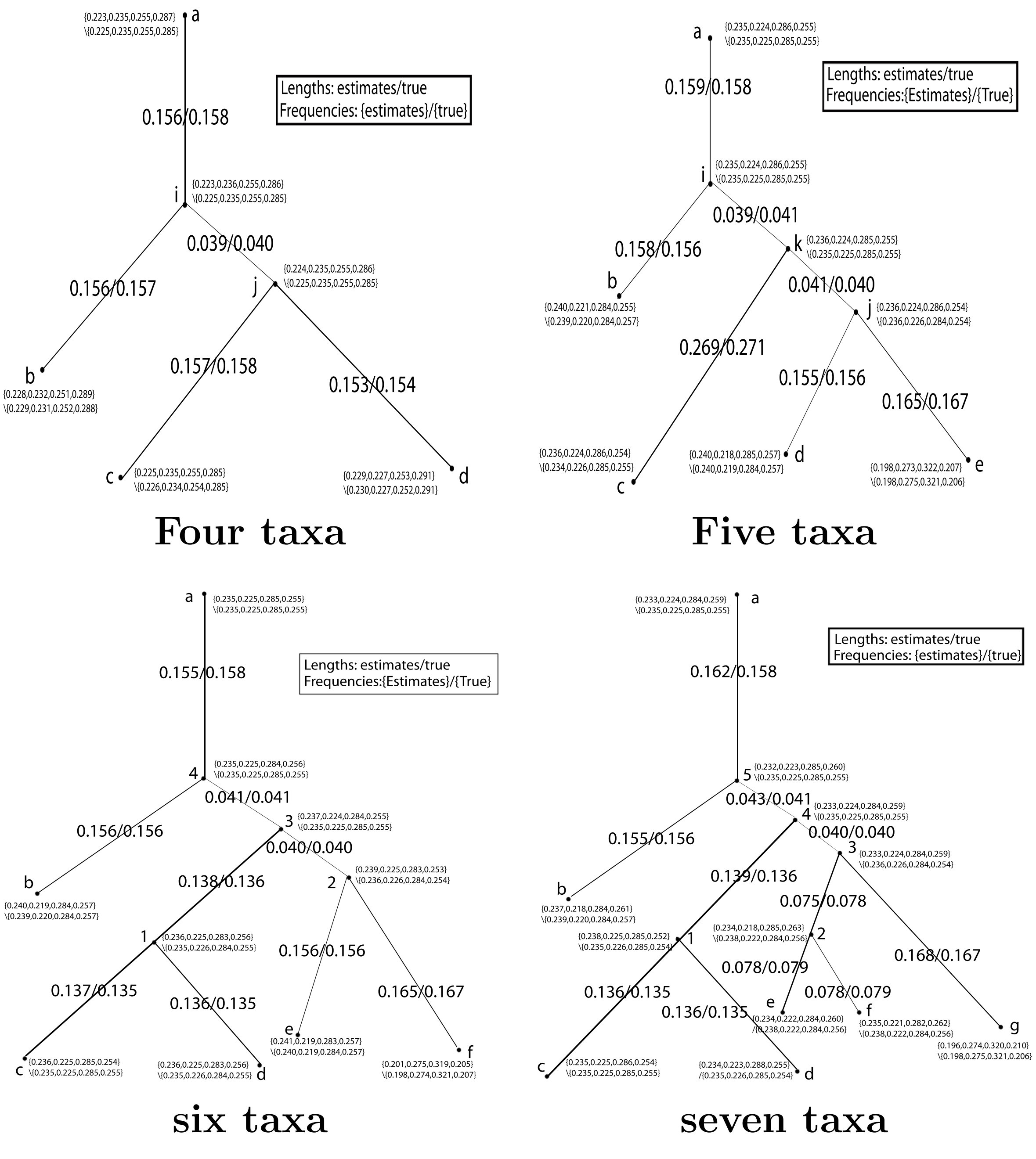
values for the first class; boxplot 5.c2 gives these differences for the second class. Boxplots 6.c1 and 6.c2 correspond to six-taxon results. From those plots, it is clear that the estimates for six taxa are closer to the true values. This result confirmed that a long edge is harder to estimate than a short edge and adding more taxa into a data set can improve the accuracy of the estimates. As a comparison, estimates in Figure 5.9a were obtained using data sets simulated under the same trees shown in Figure 5.8 and the same parameters of NSGTR models, but a different sequence length of 10,000 sites. Not surprisingly, estimation for 10,000 data sets was less precise than for 50,000 since around 75% of the possible site patterns were not observed in six-taxon with 10,000 sites data set as shown in Table 5.5. Thus the large differences between the estimates and the true values in the estimation results for six-taxa data set with 10,000 sites were likely a consequence of this information loss. Based on those observations, large samples are needed for accurate estimates.

Table 5.5: The numbers of distinct site patterns in simulated 10,000 and 200,000 sites data sets

Taxa	TRUE	10k	200k
four	256	211	256
five	1024	532	967
six	4096	889	2692
seven	16384	1230	5099

5.7.4 Mixtures and Compositional Heterogeneity

As shown in Figure 5.10, to test the hypothesis that mixtures are useful for dealing with compositional heterogeneity, we created a four-taxon tree for which the



Four taxa

Five taxa

six taxa

seven taxa

Figure 5.7: True values and estimates of combined parameters under BH2 + I model of simulated four, five, six, and seven-taxon data sets

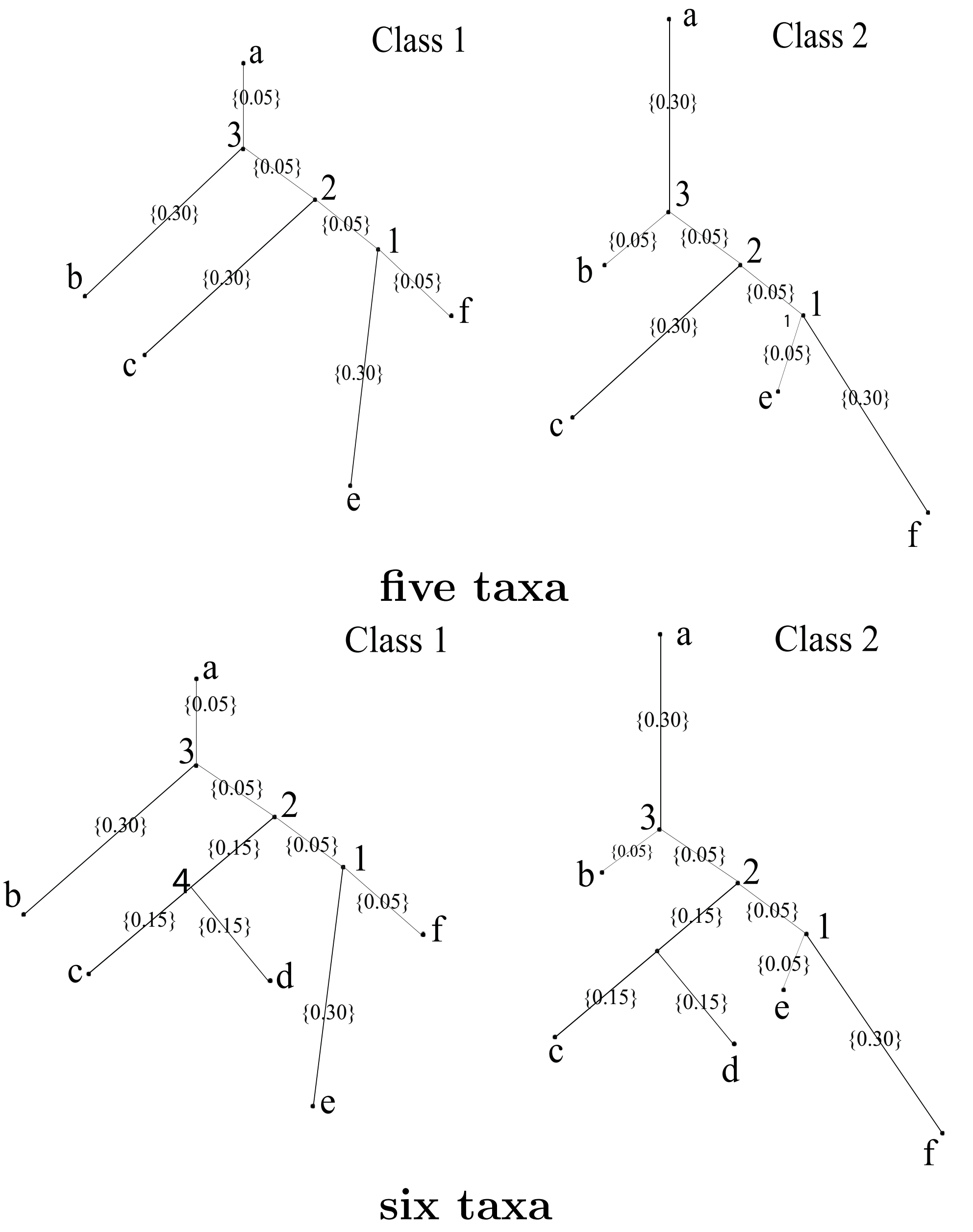


Figure 5.8: Comparing the performances of five and six-taxon trees. Note: The six-taxon tree is created by breaking the long branch (2, c) in five-taxon tree.

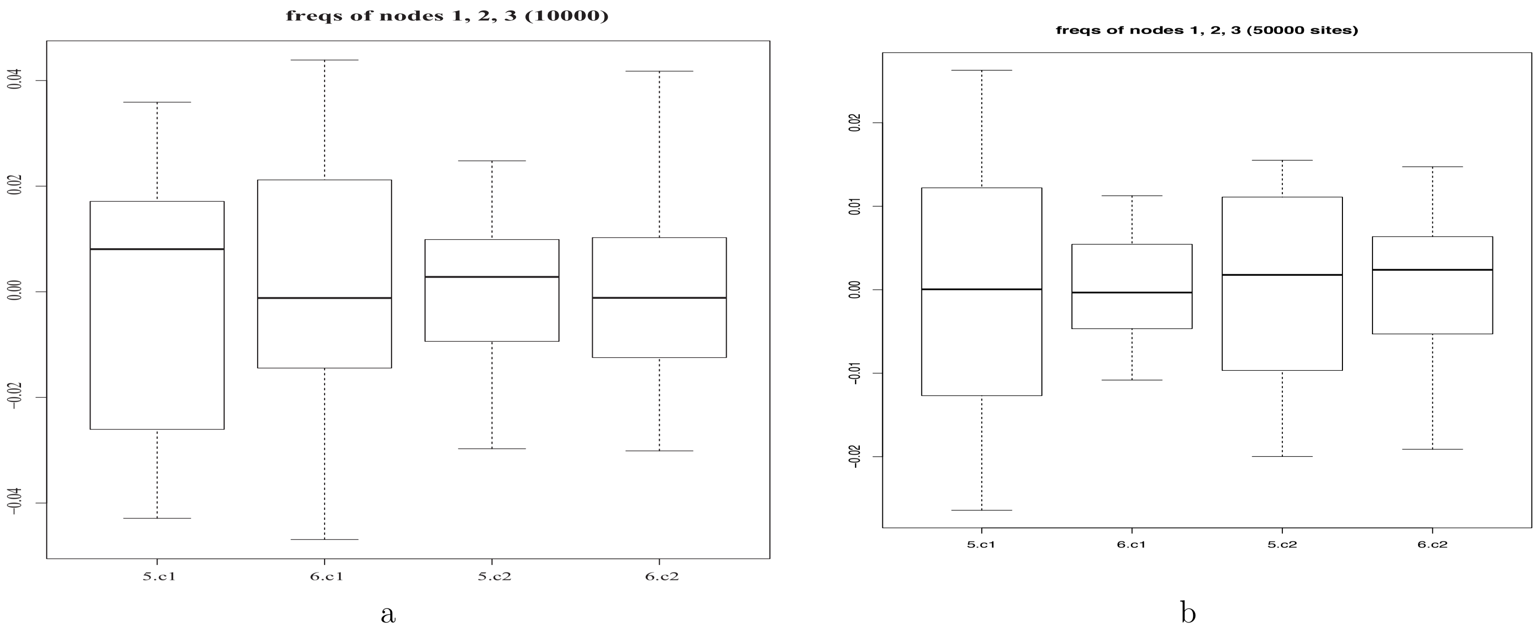


Figure 5.9: Boxplots of the differences between the estimates and true values of the frequencies of internal nodes 1, 2 and 3 in Figure 5.8. The sequence lengths in (a) and (b) are 10,000 and 50,00 sites. "5.c1" and "5.c2" are the boxplots of five-taxon estimates for the first and second classes; "6.c1" and "6.c2" are correspond to six-taxon results.

base character frequencies of neighbor leaf nodes are very different but non-neighbors had similar frequencies. From this figure, we can see that nodes a and c have very similar base character frequencies; nodes b and d also have very similar base character frequencies. When estimating, we fitted the following models: the BH2+I (the true model), BH+I, GG+Γ2, and GTR+Γ2+I models. The summary of log-likelihoods is in Table 5.6. The GG+Γ2 and GTR+Γ2+I models both succumbed to a 'base frequencies attraction' artefact whereby sequences of similar base frequency properties were grouped together; they selected the tree that grouped nodes a and c together (a, c) and nodes b and d in another group (b, d) instead of the correct groups (a, b) and (c, d). In contrast, both the BH+I and BH2+I models estimated the correct tree. This illustrates that the BH model can overcome the potential pitfall of base frequency attraction. However, the log-likelihood -848918 in the BH+I model is much smaller than the log-likelihood -848116 in the BH2+I model and the former model was significantly excluded by a likelihood ratio test (p-value = 0.0).

Table 5.6: Loglikelihoods of the dataset which has potential long/short branches attraction (four taxa)

Model	abcd	acbd	adbc	multinomial
GTR + Γ2 +I	-875275	*-874703*	-875373	-847997
GG + Γ2	-880430	*-879987*	-880594	-847997
BH + I	*-848918*	-849520	-849543	-847997
BH2 + I	*-848116*	-848478	-848501	-847997

abcd: true tree. (a, (b, (c, d)))
acbd: wrong tree. (a, (c, (b, d)))
adbc: wrong tree. (a, (d, (b, c)))
The cells format with italic and bold font are the best estimate of each model.

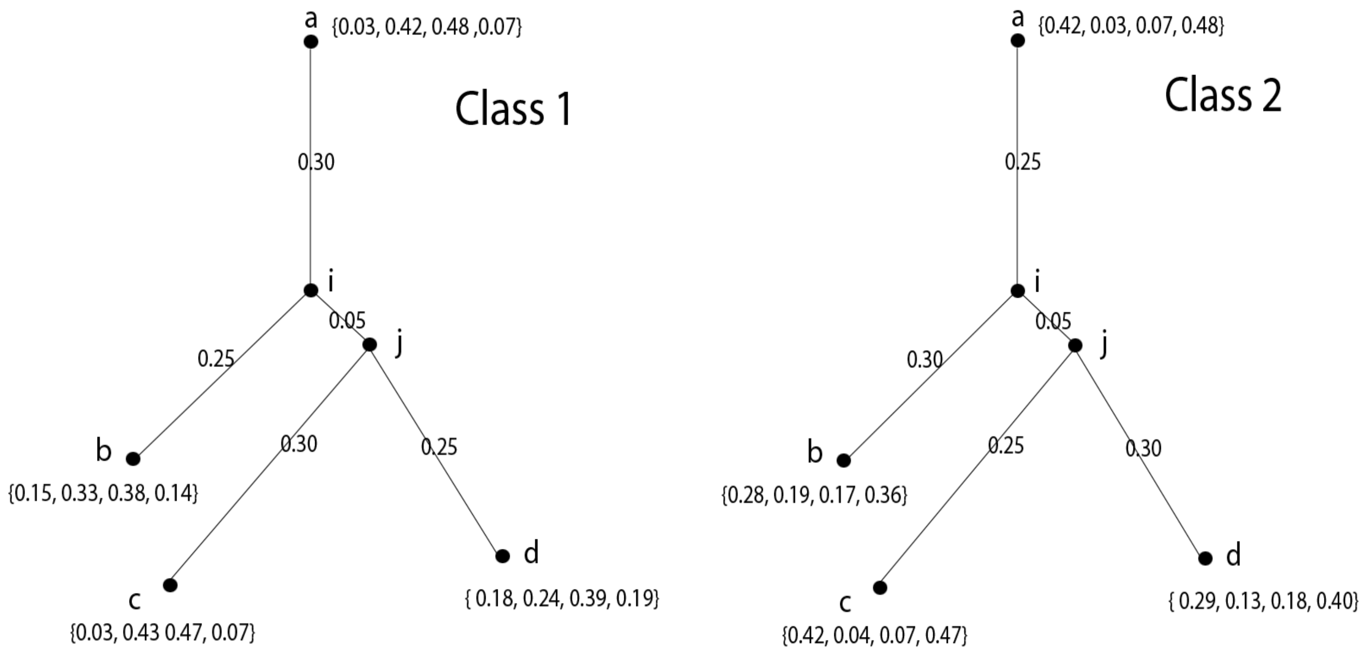


Figure 5.10: The trees in different classes with base frequency heterogeneity for neighbors

5.7.5 The Impact of the Number of Taxa on Estimation

When estimating the mixture model using ML, ideally, the number of parameters should be smaller than the number of distinct site patterns (cf. Chapter 1 in Durbin et al. 1998). Table 5.1 lists the number of possible distinct site patterns and the number of parameters in different mixture models. With only three taxa, there are 64 site patterns. A one-class mixture model (i.e., a usual BH model) has 39 parameters whereas with two classes there are 80 parameters; in the latter case, over-fitting becomes a problem. Our simulations have confirmed that it is very hard to distinguish two classes in a three-taxon data set.

Table 5.1 shows that as a function of the number of taxa, the number of parameters does not increase as fast as the number of site patterns. This suggests it will be possible to estimate mixtures with more classes when the number of taxa becomes larger. However another problem arises in this case since not all distinct site patterns will be observed from data sets of a given sequence length. As the number of distinct site patterns increases, the probability of observing some patterns will be vanishingly small. It is normal that some possible site patterns will be missing in data with six or seven taxa for reasonable-sized data sets. Table 5.5 shows the numbers of distinct patterns in simulated data sets with four to seven taxa and 10,000 or 200,000 sites. For the worst case of seven taxa, more than half of the possible patterns were not observed even with a data set of 200,000 sites; more than 90% of the patterns were missing with 10,000 sites.

5.7.6 Plasmodium *Data Set Results*

As an example with real data, we used the data set in Davalos and Perkins (2008). In this data set, there are 8 species from the malaria parasites of genus *Plasmodium* and 77313 sites after alignment gaps are removed. Among those sites, there are 3865 distinct site patterns which is around 6% of total possible sites. All estimates were obtained from the fixed topology given in Figure 1(A) of Davalos and Perkins (2008) for the following models: GTR, GG, BH, BH+I, BH2, BH2+I, BH3, BH3+I and BH4. The overall summary for the estimates is given in Table 5.7. The following likelihood ratio tests were significant: BH against GTR/GG, BH+I against BH, BH2 against BH+I, BH2+I against BH2, BH3 against BH2+I, and BH3+I against BH3. However, we found that the BH4 was not significantly better than the BH3+I model (p-value = 0.999, d.f.= 156, $\chi^2 = 104.4$). Therefore, for this real data set, BH3+I is the most appropriate model.

As shown earlier, the estimates of weighted leaf node frequencies converge to the empirical leaf node frequencies. When computing weighted frequency for the character state r at node a, $\hat{\pi}_r^a$, we used the following equation.

$$\hat{\pi}_r^a = \sum_{j=1}^{K} \sum_{s=1}^{4} \hat{F}_{(a,b)}^j(r,s)\hat{\omega}_j \tag{5.16}$$

where s is the character state at the node b which is the node directly connected with node a in the edge (a,b); $\hat{\omega}_j$ is the estimate of the weight of the j^{th} class; $\hat{F}_{(a,b)}^j(r,s)$ is the estimate of the joint probability matrix of edge (a,b) in the j^{th} class.

Table 5.7: Estimates for *Plasmodium* data set (Figure 1(A) in Davalos and Perkins (2008))

model	log likelihoods	w.inv	w.c1	w.c2	w3	w4	f.inv.a	f.inv.c	f.inv.g	f.inv.t
Empirical	-320007.1	na	na	na	na	na	0.40	0.115	0.195	0.29
GG	-341101.48	na	na	na	na	na	na	na	na	na
GTR	-348615.75	na	na	na	na	na	na	na	na	na
BH	-335974.11	na	na	na	na	na	na	na	na	na
BH + I	-327639.51	0.48	0.52	na	na	na	0.38	0.12	0.21	0.29
BH2	-326868.10	na	0.45	0.55	na	na	na	na	na	na
BH2 + I	-325868.10	0.44	0.31	0.25	na	na	0.34	0.13	0.23	0.30
BH3	-325692.9	na	0.35	0.29	0.36	na	na	na	na	na
BH3 + I	-325572.6	0.44	0.23	0.14	0.19	na	0.37	0.12	0.22	0.29
BH4	-325520.4	na	0.56	0.16	0.06	0.22	na	na	na	na

For the estimates of leaf node frequencies under BH, BH+I, BH2, BH2+I, BH3, BH3+I and BH4 models, we had the sums of squares of the differences between the empirical and estimated leaf nodes frequencies of 4.84e-02, 1.37e-08, 4.84e-02, 1.36e-08, 4.84e-02, 2.71e-08, 4.84e-02 respectively. From this, it appears that estimates under BH$(K - 1)$ + I models are, in general, closer to the empirical leaf node frequencies than the estimates under BHK models. Clearly, for this particular data set, it is important to take into account invariable sites as a distinct class.

As shown in Table 5.7, similar invariable frequencies and invariable site class weights were estimated consistently by the BH+I, BH2+I and BH3+I models. The estimates of invariable site state frequencies from all BH$(K - 1)$ + I model all converged to the corresponding empirical invariant site state frequencies. For the *Plasmodium* data set, therefore, it is likely that the proportion of invariable sites is roughly 44%.

In our analysis, we compared edge length estimates yielded by the various models, in the following manner: BH versus BH+I, BH2 versus BH2+I; BH3 versus BH3+I. We found that the estimates of edge lengths in each class under the BHK + I models are larger than the ones under BHK models. Then we weighted the edge lengths of all classes to obtain estimates of the overall edge lengths for all edges on this topology. The estimates of overall edge lengths are shown in Table 5.8. Comparing edge-by-edge, it is not possible to make general statements about the relative lengths yielded by the BH$(K - 1)$ + I models versus the BHK models. However, when edge lengths are summed over all edges, it is clear that the BH$(K - 1)$ + I

Table 5.8: Weighted expected numbers of substitutions over all classes of *Plasmodium* data set

Branch	BH	BH + I	BH2	BH2 + I	BH3	BH3 + I
61	0.124	0.155	0.099	0.168	0.184	0.204
52	0.182	0.218	0.218	0.193	0.240	0.316
43	0.018	0.019	0.024	0.025	0.019	0.024
54	0.095	0.131	0.185	0.157	0.156	0.301
65	0.030	0.022	0.015	0.012	0.024	0.178
g6	0.104	0.128	0.093	0.128	0.201	0.162
3b	0.024	0.025	0.026	0.026	0.029	0.027
4c	0.038	0.039	0.041	0.044	0.049	0.048
1f	0.010	0.009	0.010	0.010	0.009	0.008
2k	0.056	0.065	0.071	0.069	0.068	0.074
1r	0.011	0.011	0.011	0.011	0.013	0.013
2v	0.099	0.106	0.107	0.110	0.111	0.104
3y	0.028	0.030	0.031	0.033	0.037	0.034
sum	0.791	0.959	0.933	0.987	1.139	1.494

models yield a greater total were larger than the estimates under the BHK models. One explanation may be that for this particular data set with 44 % invariable sites, when using a BH$(K-1)$+I model, e.g., the BH2+I model, it does remove the invariable site effect. Thus the variable classes did get more accurate estimates and showed longer edge lengths along the tree in each classes compared with the variable classes in a BHK model, e.g., the BH2 model. The observation that the sum of weighted edge lengths for a BH$(K-1)$+I model is larger may reflect an accuracy improvement due to appropriate adjustment for invariable sites.

5.7.7 Identifiability of the Mixture BH model

In a mixture BH model, the identifiability issues exist. When considering identifiability issues, we have to consider not only the estimates in each class, but also the topologies.

From our experiments, it is clear that the estimates in each class are non-identifiable. This is not a surprise conclusion. We have shown that the estimate of a BH model is non-identifiable. As an extension of the BH model, estimates in each class are still the BH estimates. Therefore, without any knowledge, we have to estimate permutations for BH estimates in each class and then go for estimating edge lengths or interpret internal frequencies. Since the estimates in each class are independent, we can go through class by class and estimate permutations for all classes.

5.8 Conclusions

We implemented a mixture-model extension of the BH model. In contrast to the BH model, it allows evolutionary process variation across different site classes. From our simulation results, the mixture model can be recovered accurately for all of data sets with 4-8 taxa. Our current implementation cannot handle many more taxa than this. Obtaining a more efficient implementation needs to be a focus of future work. Using likelihood ratio tests, we are able to identify the correct model among different models. The mixture-model also shows promise for the case where base frequency bias exists and is accompanied by rate-across-sites variation.

A5.1. Leaf Node Frequencies

In the mixture model, the leaf frequency of character r at leaf node a in an edge (a,l) can be calculated by summing over all classes and the character state s of internal node l in the joint probability matrix $F^j_{(a,l)}(r,s)$, where $j \in \{1,\ldots,K\}$, or $\pi^a_r = \sum_{j=1}^K \omega_j \sum_{s=1}^4 F^j_{(a,l)}(r,s)$.

Assume that the observed sequences have N sites. The log-likelihood is

$$l(X|F,\Omega) = \sum_{i=1}^N \log\left[\sum_{j=1}^K P(\underline{x}_i|F^j)\omega_j\right] \tag{5.17}$$

We first consider estimating ω_j. Using the method of Lagrange multipliers, the maximizer of the loglikelihood is also the maximizer of

$$\sum_{i=1}^N \log\left[\sum_{j=1}^K P(\underline{x}_i|F^j)\omega_j\right] + \lambda(1 - \sum_{l=1}^K \omega_l) \tag{5.18}$$

for some $\lambda > 0$. Taking derivative with respect to ω_j, setting the derivative to zero and solving for λ gives

$$\lambda \;\; = \;\; \sum_{i=1}^N \frac{P(\underline{x}_i|F^j)}{P(\underline{x}_i|F,\underline{\omega})} \tag{5.19}$$

where $P(\underline{x}_i|F,\underline{\omega}) = \sum_{j=1}^K P(\underline{x}_i|F^j)\omega_j$. Multiplying ω_j and summing over j gives

$$\lambda\sum_{j=1}^K \omega_j = \lambda = \sum_{j=1}^K \omega_j \sum_{i=1}^N \frac{P(\underline{x}_i|F^j)}{P(\underline{x}_i|F,\underline{\omega})} = N \tag{5.20}$$

Since $\lambda\omega_j = \sum_{i=1}^{N} \frac{P(\underline{x}_i|F^j)\omega_j}{P(\underline{x}_i|F,\underline{\omega})} = \sum_{i=1}^{N} \frac{P(S_i=j|\underline{x}_i,F)P(\underline{x}_i|F,\underline{\omega})}{P(\underline{x}_i|F,\underline{\omega})} = \sum_{i=1}^{N} P(S_i = j|\underline{x}_i, F)$, the updating equation for ω_j is

$$\omega_j = \frac{1}{N}\sum_{i=1}^{N} P(S_i = j|\underline{x}_i, F, \underline{\omega}) \tag{5.21}$$

Consider the edge (a, l) in the figure 5.11. The node a is the leaf node; the node l is the internal node and it is connected with a subtree T_l. Let the conditional probability of the data in the subtree T_l, X_{T_l}, given the state s at node l be $P^j_{l,T_l}(X_{T_l}|X_l = s)$ for the j^{th} class. The loglikelihood of N distinct sites is

$$l(X|F,\underline{\omega}) = \sum_{i=1}^{N} \log\left[\sum_{j=1}^{K}\sum_{s=1}^{4} F^j_{(a,l)}(r_i, s)P^j_{l,T_l}(X_{T_l}|X_l = s, F^j)\omega_j\right] \tag{5.22}$$

where r_i is the character state at node a in site i. We now consider ML estimation of $F^j_{(a,l)}(r, s)$. Using the method of Lagrange multipliers, the maximizer of the log likelihood is also the maximizer of

$$\sum_{i=1}^{N} \log\left[\sum_{j=1}^{K}\sum_{s=1}^{4} F^j_{(a,l)}(r_i, s)P^j_{l,T_l}(X_{T_l}|X_l = s, F^j)\omega_j\right] + \lambda(1 - \sum_{r=1}^{4}\sum_{s=1}^{4} F^j_{(a,l)}(r, s)) \tag{5.23}$$

for some $\lambda > 0$. Taking derivative with respect to $F^j_{(a,l)}(r, s)$, setting the derivative to zero and solving for λ gives

$$\lambda = \sum_{i=1}^{N} \mathbf{1}\{r_i = r\} \frac{P^j_{l,T_l}(X_{T_l}|X_l = s)\omega_j}{P(\underline{x}_i|F,\underline{\omega})} \tag{5.24}$$

Multiplying by $F^j_{(a,l)}(r,s)$ and summing over r, s gives

$$\lambda \sum_{r=1}^{4}\sum_{s=1}^{4} F^j_{(a,l)}(r,s) = \sum_{i=1}^{N}\sum_{r=1}^{4} \mathbf{1}\{r_i = r\} \frac{\sum_{s=1}^{4} F^j_{(a,l)}(r,s) P^j_{l,T_l}(X_{T_l}|X_l = s)\omega_j}{P(\underline{x}_i|F,\underline{\omega})}$$

Given

$$\sum_{r=1}^{4}\sum_{r=1}^{4} F^j_{(a,l)}(r,s) = 1$$

and

$$\sum_{s=1}^{4} F^j_{(a,l)}(r,s) P^j_{l,T_l}(X_{T_l}|X_l = s)\omega_j = P(\underline{x}_i, S_i = j|F,\underline{\omega}),$$

λ can be solved as

$$\lambda \;=\; \sum_{i=1}^{N} \frac{P(\underline{x}_i, S_i = j|F,\underline{\omega})}{P(\underline{x}_i|F,\underline{\omega})} = \sum_{i=1}^{N} P(S_i = j|\underline{x}_i, F, \underline{\omega}) = N\omega_j \tag{5.25}$$

Recall that we took derivative with respect to $F^j_{(a,l)}(r,s)$ and set this derivative as zero when maximizing the likelihood in (5.22). Reorganizing this derivative and multiplying by $F^j_{(a,l)}(r,s)$ gives

$$F^j_{(a,l)}(r,s) = \frac{1}{\lambda} \sum_{i=1}^{N} \mathbf{1}\{r_i = r\} \frac{F^j_{(a,l)}(r,s) P^j_{l,T_l}(X_{T_l}|X_l = s, F^j)\omega_j}{P(\underline{x}_l|F,\underline{\omega})} \tag{5.26}$$

Replacing λ by the result in (5.25) gives

$$F^j_{(a,l)}(r,s) = \frac{1}{N\omega_j}\sum_{i=1}^{N}\mathbf{1}\{r_i = r\}\frac{F^j_{(a,l)}(r,s)P^j_{l,T_l}(X_{T_l}|X_l = s, F^j)\omega_j}{P(\underline{x}_l|F,\underline{\omega})} \tag{5.27}$$

Thus, $\sum_{j=1}^{K}\omega_j\sum_{s=1}^{4}F^j_{(a,l)}(r,s)$ satisfies

$$\begin{aligned}\sum_{j=1}^{K}\omega_j\sum_{s=1}^{4}F^j_{(a,l)}(r,s) &= \sum_{i=1}^{N}\sum_{j=1}^{K}\omega_j\sum_{s=1}^{4}\frac{1}{N\omega_j}\mathbf{1}\{r_i = r\}\frac{F^j_{(a,l)}(r,s)P^j_{l,T_l}(X_{T_l}|X_l = s, F^j)\omega_j}{P(\underline{x}_i|F,\underline{\omega})}\\ &= \frac{1}{N}\sum_{i=1}^{N}\mathbf{1}\{r_i = r\}\frac{\sum_{j=1}^{K}\sum_{s=1}^{4}F^j_{(a,l)}(r,s)P^j_{l,T_l}(X_{T_l}|X_l = s, F^j)\omega_j}{P(\underline{x}_i|F,\underline{\omega})}\\ &= \frac{1}{N}\sum_{i=1}^{N}\mathbf{1}\{r_i = r\}\frac{P(\underline{x}_i|F,\underline{\omega})}{P(\underline{x}_i|F,\underline{\omega})}\\ &= \frac{1}{N}\sum_{i=1}^{N}\mathbf{1}\{r_i = r\}\end{aligned} \tag{5.28}$$

$\sum_{i=1}^{N}\mathbf{1}\{r_i = r\}$ is the number of the character r appears in node a, denoted by n^a_r.

The marginal frequency of the character state r at node a is

$$\sum_{j=1}^{K}\omega_j\sum_{s=1}^{4}F^j_{(a,l)}(r,s) = \frac{n^a_r}{N} \tag{5.29}$$

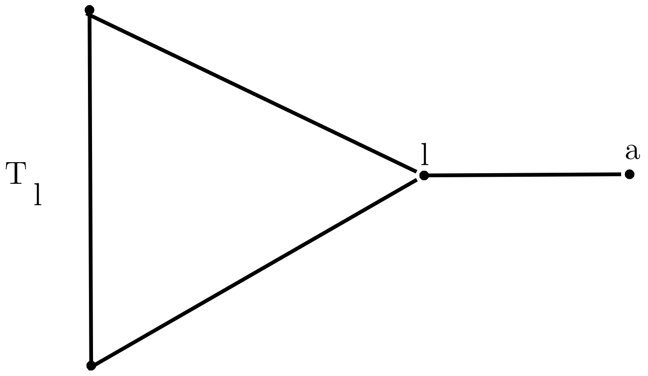


Figure 5.11: Illustration of branch (a, l) and subtree T_l

Chapter 6

Discussion

6.1 The Best Fit Binning Strategy

One of the issues that deserves additional attention in our rate matrix construction procedure is the determination of an appropriate binning strategy. Based on the data extracted from the PANDIT database, we proposed a binning strategy with 4 bins of equal size. It is expected, however, that different distributions of pairwise similarities will give different binning strategies. It is also true that the distributions of pairwise similarities in other data sets are rarely the same as the distribution of pairwise similarities obtained from the data in our experiment.

Our current results showed that an equal size strategy may do better than an equal width one if the distribution of pairwise similarities is far from uniform. It may therefore be promising to investigate through simulation the performances of equal size versus width strategies as the distribution of pairwise similarities gradually changes from extremely skewed to symmetric, and then to uniform. Since we would know the true model used in such simulations, we could calculate the aveMSEs and aveBiases for all simulations and use those as measures to determine the relationship between binning strategies and distributions. It is possible that the binning strategies that we found to be optimal for the PANDIT database are not the best for some analyses. Cross-validation methods may be provide a way of testing the performance of a binning strategy for a given data set.

Rates-across-sites variation is widely recognized as a prevalent phenomena and adjusting for it has become standard practice in phylogenetic analysis. Amongst empirical amino acid rate matrices, the LG matrix is the only one which was constructed adjusting for rates-across-sites variation. Incorporating rates-across-sites variation into our binning approach would be of substantial value.

6.2 Fitting a more Realistic BH Mixture Model

The BH mixture model is useful for some of the cases that we have examined in Chapter 5. Using the results from simulations and the analyses of the *Plasmodium* data set of Chapter 4 as a motivation, we could develop a method which may dynamically find the most appropriate mixture model for any particular data set under investigation.

In recent work by Jayaswal et al. (2011), two simplified BH models were proposed. The SBH model in Jayaswal et al. (2011) assumes that the processes are globally stationary, nonreversible and nonhomogeneous; the RBH model assumes, in addition to the SBH model assumptions, that the processes along edges are reversible. These two simplified BH models both make global assumptions and may not identify the local differences in evolutionary processes. For instance, recall that the clusters in the tree of the *Plamsodium* data in Chapter 4: the cluster made up of nodes, 2, k and v and the cluster of all other nodes. Within each of these clusters the recovered NSGTR models are very similar. These observations indicate that in real data, closely related lineages may display similar evolutionary dynamics and thus a single model may be reasonable within suitably defined clusters. The situation is illustrated in Figure 6.1 where subtree 1 may be modeled by a stationary or stationary and reversible model; subtree 2 need not have those constraints.

Determining how to cluster species dynamically will be a challenge for this method. Ababneh et al. (2006) discussed several methods of testing for violations of homogeneous conditions. Alternatively, starting from an unconstrainted model, one might successively add to a cluster those edges that gave the smallest decrease in likelihood.

6.3 Application Issues in ML Estimation

In much of this thesis, we have used ML estimation of parameters. Two significant recurring difficulties were multiple local maxima and computational cost. To find global maxima, we used randomly generated initial values. This strategy usually works but is expensive since, in some cases, hundreds of optimization processes are needed to find the globally largest likelihood. One way to reduce computational cost is to add more constraints on parameters. Adding constraints reduces the size of the parameter space and thus should make optimization easier. As an example, we know that at the optimum, the leaf node frequencies equal empirical leaf node frequencies. We could use this as a constraint in optimization and as a constraint on the initial values of parameters.

Improving the quality of software may help to improve efficiency too. The current application was constructed under the JNI+C+NAG library structure. The Java language was used because it is easier to implement than with C. However the Java virtual machine and JNI require a lot of CPU and memory. The optimization routine used in this thesis is the e04ucf routine in the NAG library, a sophisticated optimization routine. Our experiments indicate that it is stable and does a good job of finding the maxima for our complicated models. However, it is worth considering whether other choices of generic routines may be able to provide the similar or more accurate results for this particular model.

Parallel computing is a method that allows an application to use multiple computing resources such as CPUs and memories at the same time. The advantage of using parallel computing mechanism is that it allows several similar jobs to be processed at the same time using different resources. In our application, the likelihood calculation for different sites could be implemented using a parallel computing algorithm. Also, if available, using a parallelized version of the optimization routines would reduce the time needed for the optimization procedures.

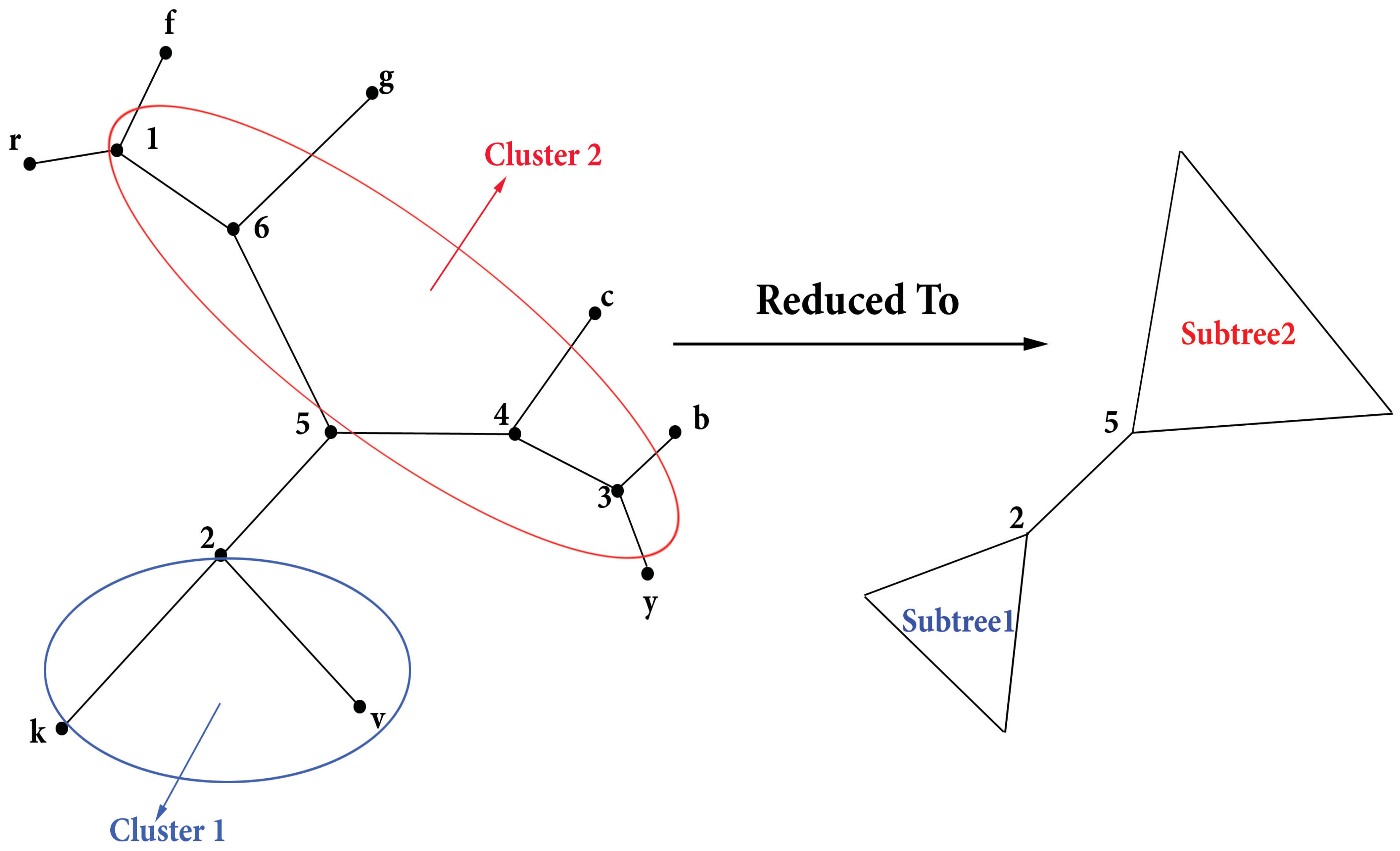


Figure 6.1: Illustration of clustering species

Bibliography

Ababneh, F., L. S. Jermiin, C. Ma, and J. Robinson. 2006. Matched-pairs tests of homogeneity with applications to homologous nucleotide sequences. Bioinformatics 22:1225–1231.

Adachi, J. and M. Hasegawa. 1996. Model of amino acid substitution in proteins encoded by mitochondrial DNA. Journal of Molecular Evolution 42:459–468.

Adachi, J., P. J. Waddell, W. Martin, and M. Hasegawa. 2000. Plastid genome phylogeny and a model of amino acid substitution for proteins encoded by chloroplast DNA. Journal of Molecular Evolution 50:348–358.

Allman, E., C. Ane, and J. A. Rhodes. 2008. Identifiability of a Markovian model of molecular evolution with Gamma-distributed rates. Advances in Applied Probability 40:229.

Arvestad, L. and W. J. Bruno. 1997. Estimation of reversible substitution matrices from multiple pairs of sequences. Journal of Molecular Evolution 45:696–703.

Barry, D. and J. A. Hartigan. 1987. Statistical analysis of Hominoid molecular evolution. Statistical Science 2:191–207.

Bateman, A., L. Coin, R. Durbin, R. Finn, V. Hollich, S. Griffiths-Jones, A. Khanna, M. Marshall, S. Moxon, E. Sonnhammer, Y. C. Studholme, D.J., and S. Eddy. 2004. The Pfam protein families database. Nucleic Acids Research 32:D138–D141.

Bernardi, G. 1993. The vertebrate genome: isochores and evolution. Molecular Biology and Evolution 10:186.

Bickel, P. J. and K. A. Doksum. 2007. Mathematical statistics : basic ideas and selected topics VOL. I. Upper Saddle River, NJ : Prentice Hall.

Boeckmann, B., A. Bairoch, R. Apweiler, M. C. Blatter, A. Estreicher, E. Gasteiger, M. J. Martin, K. Michoud, C. O'Donovan, I. Phan, S. Pilbout, and M. Schneider. 2003. The SWISS-PROT protein knowledgebase and its supplement TrEMBL in 2003. Nucleic Acids Research 31:365–370.

Boussau, B., S. Blanquart, A. Necsulea, N. Lartillot, and M. Gouy. 2008. Parallel adaptations to high temperatures in the Archaean eon. Nature 456:942–945.

Boussau, B. and M. Gouy. 2006. Efficient likelihood computations with nonreversible models of evolution. Systematic Biology 55:756–768.

Bromham, L. 2009. Why do species vary in their rate of molecular evolution? Biology Letters 5:401–404.

Chai, J. and E. A. Housworth. 2011. On Rogers' proof of identifiability for the GTR + Γ + I model. Systematic Biology 60:713–718.

Chang, J. 1996. Full reconstruction of Markov models on evolutionary trees: identifiability and consistency. Mathematical Biosciences Pages 51–73.

Chen, F., A. J. Mackey, C. J. Stoeckert Jr, and D. S. Roos. 2006. OrthoMCL-DB: querying a comprehensive multi-species collection of ortholog groups. Nucleic Acids Research 34:D363–8.

Chenna, R., H. Sugawara, T. Koike, R. Lopez, T. J. Gibson, D. G. Higgins, and J. D. Thompson. 2003. Multiple sequence alignment with the Clustal series of programs. Nucleic Acids Research 31:3497–3500.

Davalos, L. M. and S. L. Perkins. 2008. Saturation and base composition bias explain phylogenomic conflict in Plasmodium. Genomics 91:433–442.

Dayhoff, M. O., R. Schwartz, and B. Orcutt. 1979. A model of evolutionary change in proteins Pages 345–352. Atlas of Protein Sequence and Struction National Biomedical Research Foundation, Washington, D.C.

Dempster, A. P., N. M. Laird, and D. B. Rubin. 1977. Maximum likelihood from incomplete data via the EM algorithm. Journal of the Royal Statistical Society.Series B (Methodological) 39:1–38.

Durbin, R., S. Eddy, A. Krogh, and G. Mitchison. 1998. Biological sequence analysis. Cambridge University Press.

Edwards, A. W. F. and L. L. Cavalli-Sforza. 1963. The reconstruction of evolution. Annals of Human Genetics 27:105–106.

Felsenstein, J. 1989. PHYLIP - phylogeny inference package (Version 3.2). Cladistics 5:164–166.

Felsenstein, J. 2004. Inferring phylogenies. Sinauer Associates Inc. :Sunderland Massachuetts.

Fitch, W. M. and E. Margoliash. 1967. A method for estimating the number of invariant amino acid coding positions in a gene using cytochrome c as a model case. Biochemical Genetics 1:65–71.

Fletcher, W. and Z. Yang. 2009. INDELible: a flexible simulator of biological sequence evolution. Molecular Biology and Evolution 26:1879–1888.

Foster, P. G. 2004. Modeling compositional heterogeneity. Systematic Biology 53:485–495.

Foster, P. G. and D. A. Hickey. 1999. Compositional bias may affect both DNA-based and protein-based phylogenetic reconstructions. Journal of Molecular Evolution 48:284–290.

Foster, P. G., L. S. Jermiin, and D. A. Hickey. 1997. Nucleotide composition bias affects amino acid content in proteins coded by animal mitochondria. Journal of Molecular Evolution 44:282–288.

Fukushima, M. 1980. Dirichlet forms and Markov processes. North Holland, New York.

Galtier, N. 2001. Maximum-likelihood phylogenetic analysis under a covarion-like model. Molecular Biology and Evolution 18:866–873.

Galtier, N. and M. Gouy. 1995. Inferring phylogenies from DNA sequences of unequal base compositions. Proceedings of the National Academy of Sciences of the United States of America 92:11317–11321.

Galtier, N. and M. Gouy. 1998. Inferring pattern and process: maximum-likelihood implementation of a nonhomogeneous model of DNA sequence evolution for phylogenetic analysis. Molecular Biology and Evolution 15:871–879.

Graur, D. and W. H. Li. 2000. Fundamentals of Molecular Evolution. Second Edition. Sinauer Associates Inc., Publishers:Sunderland, Massachusetts.

Gu, X., Y. X. Fu, and W. H. Li. 1995. Maximum likelihood estimation of the heterogeneity of substitution rate among nucleotide sites. Molecular Biology and Evolution 12:546–57.

Guindon, S. and O. Gascuel. 2003. A simple, fast, and accurate algorithm to estimate large phylogenies by maximum likelihood. Systematic Biology 52:696–704.

Hasegawa, M. and T. Hashimoto. 1993. Ribosomal RNA trees misleading? Nature 361:23.

Hasegawa, M., H. Kishino, and T. Yano. 1985. Dating of the Human-ape splitting by a molecular clock of mitochondrial DNA. Journal of Molecular Evolution 22:160–174.

Hasegawa, M., H. Kishino, and T. Yano. 1987. Man's place in Hominoidea as inferred from molecular clocks of DNA. Journal of Molecular Evolution 26:132–147.

Henikoff, S. and J. G. Henikoff. 1992. Amino acid substitution matrices from protein blocks. Proceedings of the National Academy of Sciences of the United States of America 89:10915–10919.

Huelsenbeck, J. P. 2002. Testing a covariotide model of DNA substitution. Molecular Biology and Evolution 19:698–707.

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

Jayaswal, V., L. S. Jermiin, L. Poladian, and J. Robinson. 2011. Two stationary nonhomogeneous Markov models of nucleotide sequence evolution. Systematic Biology 60:74–86.

Jayaswal, V., J. Robinson, and L. Jermiin. 2007. Estimation of phylogeny and invariant sites under the general Markov model of nucleotide sequence evolution. Systematic Biology 56:155–162.

Jin, L. and M. Nei. 1990. Limitations of the evolutionary parsimony method of phylogenetic analysis. Molecular Biology and Evolution 7:82–102.

Jones, D. T., W. R. Taylor, and J. M. Thornton. 1992. The rapid generation of mutation data matrices from protein sequences. Computer Applications in the Biosciences : CABIOS 8:275–282.

Jukes, T. H. 1986. Silent nucleotide substitutions and G+C content of some mitochondrial and bacterial genes. Journal of Molecular Evolution 24:39.

Kishino, H., T. Miyata, and M. Hasegawa. 1990. Maximum likelihood inference of protein phylogeny and the origin of Chloroplasts. Journal of Molecular Evolution 31:151–160.

Kosiol, C. and N. Goldman. 2005. Different versions of the Dayhoff rate matrix. Molecular Biology and Evolution 22:193–199.

Kuhner, M. K. and J. Felsenstein. 1994. A simulation comparison of phylogeny algorithms under equal and unequal evolutionary rates. Molecular Biology and Evolution 11:459.

Le, S. Q. and O. Gascuel. 2008. An improved general amino acid replacement matrix. Molecular Biology and Evolution 25:1307–1320.

Lockhart, P. J., C. J. Howe, D. A. Bryant, T. J. Beanland, and A. W. Larkum. 1992. Substitutional bias confounds inference of cyanelle origins from sequence data. Journal of Molecular Evolution 34:153–162.

Lockhart, P. J., A. W. Larkum, M. Steel, P. J. Waddell, and D. Penny. 1996. Evolution of chlorophyll and bacteriochlorophyll: the problem of invariant sites in sequence analysis. Proceedings of the National Academy of Sciences of the United States of America 93:1930–1934.

Lockhart, P. J., M. A. Steel, M. D. Hendy, and D. Penny. 1994. Recovering evolutionary trees under a more realistic model of sequence. Molecular Biology and Evolution 11:605–612.

Lopez, P., D. Casane, and H. Philippe. 2002. Heterotachy, an important process of protein evolution. Molecular Biology and Evolution 19:1–7.

McLachlan, G. J. and T. Krishnan. 1997. The EM algorithm and extensions. Wiley, New York.

Minin, V. N. and M. A. Suchard. 2008a. Counting labeled transitions in continuous-time Markov models of evolution. Journal of Mathematical Biology 56:391–412.

Minin, V. N. and M. A. Suchard. 2008b. Fast, accurate and simulation-free stochastic mapping. Philosophical Transactions of the Royal Society of London.Series B, Biological sciences 363:3985–3995.

Montero, L. M. 1990. Gene distribution and isochore organization in the nuclear genome of plants. Nucleic Acids Symposium Series 18:1859.

Mooers, A. and E. C. Holmes. 2000. The evolution of base composition and phylogenetic inference. Trends in Ecology and Evolution (Personal edition) 15:365–369.

Mount, D. W. 2004. Bioinformatics : sequence and genome analysis. Cold Spring Harbor, N.Y. : Cold Spring Harbor Laboratory Press.

Müller, T., R. Spang, and M. Vingron. 2002. Estimating amino acid substitution models: a comparison of Dayhoff's estimator, the resolvent approach and a maximum likelihood method. Molecular Biology and Evolution 19:8–13.

Müller, T. and M. Vingron. 2000. Modeling amino acid replacement. Journal of Computational Biology 7:761–776.

Nei, M., R. Chakraborty, and P. A. Fuerst. 1976. Infinite allele model with varying mutation rate. Proceedings of the National Academy of Sciences of the United States of America 73:4164–4168.

Neyman, J. 1971. Molecular studies of evolution: a source of novel statistical problems Pages 1–27. Statistical Decision Theory and Related Topics Academic Press, New York.

Olsen, G. J. 1987. Earliest phylogenetic branchings: comparing rRNA-based evolutionary trees inferred with various techniques. Cold Spring Harbor Symposia on Quantitative Biology 52:825–837.

Oscamou, M., D. McDonald, V. B. Yap, G. A. Huttley, M. E. Lladser, and R. Knight. 2008. Comparison of methods for estimating the nucleotide substitution matrix. BMC Bioinformatics 9:511.

Rambaut, A. and N. C. Grassly. 1997. Seq-Gen: an application for the Monte Carlo simulation of DNA sequence evolution along phylogenetic trees. Computer Applications in the Biosciences : CABIOS 13:235–238.

Rogers, J. S. 2001. Maximum likelihood estimation of phylogenetic trees is consistent when substitution rates vary according to the invariable sites plus gamma distribution. Systematic Biology 50:713.

Saitou, N. and M. Nei. 1987. The neighbor-joining method: a new method for reconstructing phylogenetic trees. Molecular Biology and Evolution 4:406–425.

Schmidt, H., K. Strimmer, M. Vingron, and A. von Haeseler. 2002. TREE-PUZZLE: maximum likelihood phylogenetic analysis using quartets and parallel computing. Bioinformatics 18:502–504.

Semple, C. A. M. and M. S. Taylor. 2009. The structure of change. Science 323:347.

Sheffield, N. C., H. Song, S. L. Cameron, and M. F. Whiting. 2009. Nonstationary evolution and compositional heterogeneity in Beetle mitochondrial phylogenomics. Systematic Biology 58:381.

Shoemaker, J. S. and W. M. Fitch. 1989. Evidence from nuclear sequences that invariable sites should be considered when sequence divergence is calculated. Molecular Biology and Evolution 6:270–289.

Song, H. 2010. When phylogenetic assumptions are violated: base compositional heterogeneity and among-site rate variation in Beetle mitochondrial phylogenomics. Systematic Entomology 35:429.

Stamatakis, A. 2006. RAxML-VI-HPC: maximum likelihood-based phylogenetic analyses with thousands of taxa and mixed models. Bioinformatics 22:2688–2690.

Steel, M. 1994. Recovering a tree from the leaf colourations it generates under a Markov model. Applied Mathematics Letters 7:19–23.

Steel, M., D. Huson, and P. J. Lockhart. 2000. Invariable sites models and their use in phylogeny reconstruction. Systematic Biology 49:225–232.

Susko, E., C. Field, C. Blouin, and A. J. Roger. 2003. Estimation of rates-across-sites distributions in phylogenetic substitution models. Systematic Biology 52:594–603.

Tamura, K. 1992. Estimation of the number of nucleotide substitutions when there are strong transition-transversion and G+C-content biases. Molecular Biology and Evolution 9:678–687.

Tavaré, S. 1986. Some probabilistic and statistical problems in the analysis of DNA sequences. Lectures on Mathematics in the Life Sciences.

Thomas, J. A., J. J. Welch, M. Woolfit, and L. Bromham. 2006. There is no universal molecular clock for invertebrates, but rate variation does not scale with body size. Proceedings of the National Academy of Sciences of the United States of America 103:7366.

Thompson, J. D., D. G. Higgins, and T. J. Gibson. 1994. CLUSTAL W: improving the sensitivity of progressive multiple sequence alignment through sequence weighting, position-specific gap penalties and weight matrix choice. Nucleic Acids Research 22:4673–4680.

Tuffley, C. and M. Steel. 1998. Modeling the covarion hypothesis of nucleotide substitution. Mathematical Biosciences 147:63–91.

Uzzell, T. and K. W. Orbin. 1971. Fitting discrete probability distributions to evolutionary events. Science 172:1089–1096.

Veerassamy, S., A. Smith, and E. R. Tillier. 2003. A transition probability model for amino acid substitutions from blocks. Journal of Computational Biology 10:997–1010.

Waddell, P. J. and M. A. Steel. 1997. General time-reversible distances with unequal rates across sites: mixing Γ and inverse Gaussian distributions with invariant sites. Molecular Phylogenetics and Evolution 8:398–414.

Wakeley, J. 1994. Substitution-rate variation among sites and the estimation of transition bias. Molecular Biology and Evolution 11:436.

Whelan, S., P. I. de Bakker, E. Quevillon, N. Rodriguez, and N. Goldman. 2006. PANDIT: an evolution-centric database of protein and associated nucleotide domains with inferred trees. Nucleic Acids Research 34:327–31.

Whelan, S. and N. Goldman. 2001. A general empirical model of protein evolution derived from multiple protein families using a maximum-likelihood approach. Molecular Biology and Evolution 18:691–699.

Wilbur, W. J. 1985. On the PAM matrix model of protein evolution. Molecular Biology and Evolution 2:434–447.

Wilgenbusch, J. C. 2003. Inferring evolutionary trees with PAUP*. Current Protocols in Bioinformatics Chapter 6:Unit 6.4.

Wu, J. and E. Susko. 2010. A test for heterotachy using multiple pairs of sequences. Molecular Biology and Evolution 263:587–589.

Yang, Z. 1993. Maximum-likelihood estimation of phylogeny from DNA sequences when substitution rates differ over sites. Molecular Biology and Evolution 10:1396.

Yang, Z. 1994a. Estimating the pattern of nucleotide substitution. Journal of Molecular Evolution 39:105–111.

Yang, Z. 1994b. Maximum likelihood phylogenetic estimation from DNA sequences with variable rates over sites: approximate methods. Journal of Molecular Evolution 39:306–314.

Yang, Z. 1995. Evaluation of several methods for estimating phylogenetic trees when substitution rates differ over nucleotide sites. Journal of Molecular Evolution 40:689.

Yang, Z. 1996. Among-site rate variation and its impact on phylogenetic analyses. Trends in Ecology Evolution 11:367.

Yang, Z. 2006. Computational molecular evolution. Oxford University Press Inc., New York.

Yang, Z. 2007. PAML 4: phylogenetic analysis by maximum likelihood. Molecular Biology and Evolution 24:1586–1591.

Yang, Z., R. Nielsen, and M. Hasegawa. 1998. Models of amino acid substitution and applications to mitochondrial protein evolution. Molecular Biology and Evolution 15:1600–1611.

Yang, Z. and D. Roberts. 1995. On the use of nucleic acid sequences to infer early branchings in the tree of life. Molecular Biology and Evolution 12:451–458.

Zou, L. 2005. Constructing a rate matrix for amino acid evolution from pairs of sequences. Master's thesis.

Zou, L., C. Field, E. Susko, and A. J. Roger. 2011a. The parameters of the Barry and Hartigan general Markov model are statistically non-identifiable. Systematic Biology doi:10.1093/sysbio/syr034 .

Zou, L., E. Susko, C. Field, and A. J. Roger. 2011b. Fitting nonstationary general-time-reversible models using the Barry-Hartigan model. in preparation .

Zwickl, D. J. and D. M. Hillis. 2002. Increased taxon sampling greatly reduces phylogenetic error. Systematic Biology 51:588–598.