INFORMATION TO USERS

This manuscript has been reproduced from the microfilm master. UMI films the text directly from the original or copy submitted. Thus, some thesis and dissertation copies are in typewriter face, while others may be from any type of computer printer.

The quality of this reproduction is dependent upon the quality of the copy submitted. Broken or indistinct print, colored or poor quality illustrations and photographs, print bleedthrough, substandard margins, and improper alignment can adversely affect reproduction.

In the unlikely event that the author did not send UMI a complete manuscript and there are missing pages, these will be noted. Also, if unauthorized copyright material had to be removed, a note will indicate the deletion.

Oversize materials (e.g., maps, drawings, charts) are reproduced by sectioning the original, beginning at the upper left-hand corner and continuing from left to right in equal sections with small overlaps. Each original is also photographed in one exposure and is included in reduced form at the back of the book.

Photographs included in the original manuscript have been reproduced xerographically in this copy. Higher quality 6" x 9" black and white photographic prints are available for any photographs or illustrations appearing in this copy for an additional charge. Contact UMI directly to order.

UMI

A Bell & Howell Information Company
300 North Zeeb Road, Ann Arbor MI 48106-1346 USA
313/761-4700 800/521-0600

PHYLOGENETIC TREE CONSTRUCTION USING MARKOV CHAIN MONTE CARLO

DISSERTATION

Presented in Partial Fulfillment of the Requirements for
the Degree Doctor of Philosophy in the
Graduate School of The Ohio State University

By

Shuying Li, M.A.S

* * * * *

The Ohio State University

1996

Dissertation Committee:

Professor D. K. Pearl, Adviser

Professor D. E. Critchlow

Professor H. Doss

Approved by

A handwritten signature in black ink, appearing to read "Daniel K. Pearl", written over a horizontal line.

Adviser

Graduate Program in
Biostatistics

UMI Number: 9710605

UMI Microform 9710605
Copyright 1997, by UMI Company. All rights reserved.

**This microform edition is protected against unauthorized
copying under Title 17, United States Code.**

UMI
300 North Zeeb Road
Ann Arbor, MI 48103

ABSTRACT

We describe a method based on Markov chain simulation to study the phylogenetic relationship in a group of DNA sequences. Our method produces a Markov chain whose stationary distribution is the conditional distribution of the phylogeny given the observed sequences. Under Jukes-Cantor's one-parameter molecular evolutionary model, several innovations make the algorithm we have developed work well in practice. For example, since phylogenetic information is described by a tree, we have created new diagnostics to handle this type of data structure. An important by-product of the Markov chain Monte Carlo phylogeny building technique is that it provides estimates and corresponding measures of variability for *any* aspect of the phylogeny under study.

Dedicated to Liming and Kelly

ACKNOWLEDGMENTS

I wish to thank Professor Dennis Pearl for guiding me through this dissertation, always encouraging me and always patient. I wish to thank Professor Hani Doss for his help and discussion about Markov chain Monte Carlo simulation technique. I also wish to thank Professor Douglas Critchlow whose comments on this dissertation were very helpful. I wish to thank my husband, Liming, for computer technique consulting. I wish to thank my parent for their constant encouragements. Finally, I wish to thank my husband and daughter being there through this entire ordeal. The love and support you have given me these past years are beyond words.

VITA

Sept. 13, 1961 Born - Datong, Shanxi, China

July 1983 B.S. Mathematics, Nankai University,
Tianjin, China

May 1990 M.A. Economics, Iowa University, Iowa
City, Iowa

March 1993 M.A.S. Applied Statistics, The Ohio
State University, Columbus, Ohio

FIELDS OF STUDY

Major Field: Biostatistics

TABLE OF CONTENTS

	Page
Abstract	ii
Dedication	iii
Acknowledgments	iv
Vita	v
List of Tables	viii
List of Figures	x
Chapters:	
1. Introduction and Literature	1
2. A Probability Distribution of the Phylogenetic Tree	9
2.1 Notation	9
2.2 A probability distribution of the phylogenetic tree	10
2.3 Conditional distribution of the phylogenetic tree given sequences of taxa	22
3. Metropolis Algorithm	24
3.1 Metropolis-Hastings sampler review	24
3.2 Two Metropolis-Hastings samplers on the phylogenetic tree space .	27
3.2.1 The pendant branch transition strategy	29
3.2.2 The local rearrangement transition strategy	35
3.3 Properties of the algorithms	37
3.4 Computer implementation	41

4.	How Well Does The Markov Chain Monte Carlo Method Work?	49
4.1	Comparison with true probabilities of topologies that consist of three in-group taxa and one outgroup	49
4.2	Diagnostics	55
4.2.1	Scaled regeneration quantile (SRQ) plot	56
4.2.2	Inter-value distance comparisons	57
4.3	Reliability of subtrees	58
4.4	Estimate of the true phylogeny	60
5.	Applications	62
5.1	The Bacteriophage T7 data	62
5.2	The 16 species of fish data	67
5.3	The viral sequences from HIV infected Russian children	76
6.	Discussion and Extensions	84
	Bibliography	87

LIST OF TABLES

Table	Page
1.1 Part of the data of aligned DNA sequences of nine Bacteriophage T7	4
4.1 Simulated four sequences with 200 sites for each	54
5.1 The triples distances of the topologies in Figure 5.1	64
5.2 The pairwise distances among the sequences of 16 species of fish using Kimura’s two-parameter model with transition/transversion ratio of 2	69
5.3 The proportions of pairwise differences among the sequences of 16 species of fish	70
5.4 The frequencies of the four most frequent topologies (listed in Figure 5.7) occurred in the two chains	72
5.5 The aligned HIV sequences in the V3 region from HIV infected Russian children	77
5.6 Proportions of pairwise differences among 18 HIV-1 V3 sequences from HIV infected Russian children	78

LIST OF FIGURES

Figure	Page
1.1 Examples of bifurcating tree	2
2.1 Two different patterns in $\mathcal{T}_4$	11
2.2 Diagram of nucleotide substitution in the one-parameter molecular evolutionary model	19
3.1 The pendant branch transition strategy	30
3.2 The local rearrangement transition strategy	36
3.3 Outer frames of tree x and tree y	39
4.1 Three possible topologies in $\mathcal{T}_3$	50
5.1 Four most frequent topologies occurred in the Markov chain using the Bacteriophage T7 data	63
5.2 Two SRQ plots using the Bacteriophage T7 data	65
5.3 Histograms of distances of trees from within and between two Markov chains using the Bacteriophage T7 data	66
5.4 Consensus trees for the Bacteriophage T7 data using four tree construction methods	67
5.5 Two SRQ plots using the 16 species of fish data	71
5.6 Histograms of distances of trees from within and between two Markov chains using the 16 species of fish data	71

5.7	Four most frequent topologies occurred in the Markov chain using the 16 species of fish data	73
5.8	Consensus trees for the 16 species of fish data using four tree construction methods	74
5.9	Two SRQ plots using the viral sequences data collected from the HIV infected Russian children	80
5.10	Histograms of distances of trees from within and between two Markov chains using the viral sequences data collected from the HIV infected Russian children	80
5.11	The plots of cumulative frequencies of the topologies in the two Markov chains using the viral sequences data collected from the HIV infected Russian children	82

CHAPTER 1

INTRODUCTION AND LITERATURE

Reconstruction of the evolutionary history of genes and species is one of the most important subjects in the current study of molecular evolution.

“All life forms, both extant and extinct, share a common origin, and their ancestors can be traced back to one or a few organisms that lived approximately 4 billion years ago. Consequently, all animals, plants and bacteria are related by descent to each other. Closely related organisms are descended from more recent common ancestors than are distantly related ones. Two important objectives of phylogenetic studies are (1) to reconstruct the correct genealogical ties between organisms and (2) to estimate the time of divergence between organisms since they last shared a common ancestor.”

(The paragraph above was quoted from Li and Graur 1991.) In phylogenetic studies, the evolutionary relationships among a group of organisms are illustrated by means of a *phylogenetic tree*. A phylogenetic tree is a graph composed of nodes and branches, in which any two adjacent nodes are connected by only one branch (see for example the trees in Figure 1.1). The trees in Figure 1.1 are called bifurcating trees because each node splits into two branches. The *nodes* represent the taxonomic units, and the branching pattern defines the relationships among the units in terms of descent and ancestry. The branching pattern of a tree is also called the *topology*.

The taxonomic units represented by nodes can be species, populations, individuals, or genes.

Figure 1.1: (a) Rooted tree (b) Unrooted tree for five taxa.

The nodes of a phylogenetic tree are further distinguished into *external nodes* and *internal nodes*. Nodes at the tips of the trees are called external nodes and represent the taxa under consideration. The rest of the nodes are called internal nodes and represent ancestors of external nodes. For example, nodes 1–5 in both Figure 1.1(a) and Figure 1.1(b) are external nodes and nodes 6–9 in Figure 1.1(a) and nodes 7–9 in Figure 1.1(b) are internal nodes.

Phylogenetic trees can be either *rooted* (as in Figure 1.1(a)) or *unrooted* (as in Figure 1.1(b)). In a rooted tree there exists a particular node which is called the root. (e.g. node 6 in Figure 1.1(a)) There is only one path from the root to each of the external nodes. The direction of these paths corresponds to evolutionary time and the root is the common ancestor of all taxa under study. An unrooted tree specifies

only the relationship among the taxa and does not define the evolutionary path. The total number of distinct topologies of rooted trees for n taxa is $\prod_{k=2}^n (2k - 3)$ (Felsenstein 1978a). And the number for distinct unrooted trees with n taxa is the same as the number for distinct rooted trees with $n - 1$ taxa.

Phylogenetic inferences are premised on the inheritance of ancestral characteristics and on the existence of an evolutionary history defined by changes in these characteristics. The hereditary information of all living organisms, with the exception of some viruses, is carried by *deoxyribonucleic acid (DNA)* molecules. DNA usually consists of two complementary chains twisted around each other to form a right-handed helix. Each chain is a linear polynucleotide consisting of four nucleotides, two *purines*: *Adenine (A)* and *Guanine (G)*, and two *pyrimidines*: *Thymine (T)* and *Cytosine (C)*.

Phylogenetic analysis seeks to infer the evolutionary history that is most consistent with a set of observed data.

Consider the situation in which the observed data is a set of the DNA sequences of the taxa under study. An example data set is shown in Table 1.1, which is part of the aligned DNA sequences of nine Bacteriophage T7. The data presented are the 63 sites, out of 1091 total sites, which are regarded as “informative” with respect to the phylogenetic history (see page 111 of Li and Graur 1991). In nearly every experimental situation the phylogenetic tree underlying the evolutionary history of the observed sequences is unknown. The data in Table 1.1, which appears in Hillis et al. (1992), is very unusual because it was laboratory-generated with a *known* phylogeny of nine taxa (eight in-group lineages and one outgroup lineage to root the tree). This property therefore makes this an ideal data set for testing techniques of phylogenetic tree construction (see our illustration in Chapter 5 below).

R	CCGGGCTCGGCTGCGACCCGCGCCCCACTGCTGCGGCGGGTCCTCCGGGGACGCTCGGCGC
J	CCGGGCCCTAGCCGTACACCCGCGTTCCACTGCCACGGCGGGTCCTCCGGTGGTGGCCAGCGT
K	TCGGGCCCCGCGCCGACACCCGCACTCCACTGCCATGGCGGGGCGCCGGTGGTGGCCAGCGT
L	TCAGGCCCCGACCGCACATCCGCACTTCACTGCCATAGCGGGGCGCGTGGTGGTGGCCAGCGT
M	CTGAGCCCCGCGCGTATACCCGTGCTCCATTGCCACGGCGGGTTCTCCAATAGTGTCCAGCGT
N	CCGGGTTCCGGTCACGCACTTACGCCCTGCCGCGCGACAAATCCTTCGGTGGCACCTGGCAC
O	CTGGACCCCGCGCGCACCTGCGCCCCGCTACCGCGATGAATCCTTCGGTGGCATCCGACAC
P	CCGGACCCCGGCCACGCCCCGTGCGCCCCGCTGCCGCGACGAATCCTTCGGTGGCACCCGGTAC
Q	CCGGGTCCCGGTCACGCACTTACGCCCTGCTGTGCGACAAATCCTTCGGTGGCACCCGGCAC

Table 1.1: Part of the data of aligned DNA sequences of nine Bacteriophage T7. Shown are the 63 informative sites.

Over the past several decades there have been many attempts to reconstruct evolutionary histories using DNA sequence data. The principal methodologies used can be divided into three major categories: distance-matrix, parsimony (Fitch 1971, Lake 1987), and maximum likelihood (Felsenstein 1981, Kishino, Miyata, and Hasegawa 1990).

For distance-matrix methods, pairwise distances of sequences for taxa need to be computed if the data do not exist in distance form (as in Table 1.1). The distance between two sequences is calculated according to some model of molecular evolution. The most commonly used evolutionary models are the one-parameter model (Jukes and Cantor 1969), in which mutation rates from any nucleotide to any other are all the same, and the two-parameter model (Kimura 1980), in which the rate of transition (mutation within purines or pyrimidines) may differ from the rate of transversion (mutation between Purine and Pyrimidine). Using these calculated distances, one then estimates the phylogeny using a variety of algorithms. The simplest one is UPGMA, which is an unweighted recursive pairwise grouping method (Sokal and Michener 1958). A generalization is the neighbor-joining method, which weighs

pairwise distances on the basis of their average divergence from all other external nodes (Saitou and Nei 1987). An alternative approach is to minimize a criterion which measures the disagreement between the branch lengths of the tree and the corresponding observed distances. The most popular such method is the technique of Fitch and Margoliash (1967), which uses a chi-square like criterion. The underlying assumption of most distance-matrix methods is additive distance, i.e. all pairwise distances are equal to the sum of the branch lengths that connect the respective taxa. This assumption is very restrictive. Most data sets do not seem to satisfy it.

Methods based on the principle of maximum parsimony have been by far the most widely used. In general, parsimony methods for inferring phylogenies operate by selecting trees that minimize the total tree length: the number of evolutionary “steps” required to explain a given set of data. A drawback of the method is that it does not explicitly count multiple mutations over a branch. Thus, the method tends to work best when the mutation rate is low and may fail badly when the mutation rate is high.

To apply a maximum likelihood approach, a concrete model of the evolutionary process that converts one sequence into another must be specified. A maximum likelihood approach to phylogenetic inference evaluates the likelihood that the given evolutionary model will yield the observed sequences; the phylogenies with the highest likelihood are the inferred phylogenies. Although the given evolutionary model can be made arbitrarily complex, the principal computer programs in current use are based on the one-parameter or two-parameter models. Even so, these programs, for example DNAMLK in PHYLIP (Felsenstein 1993), cannot guarantee finding a global maximum of the likelihood. Moreover, no diagnostics are given to judge if the

estimates produced are indeed at the global maximum.

More comprehensive reviews of phylogenetic construction methods may be found in Swofford et al. (1996), and Hillis, Marble, and Moritz (1996). All of these methods produce only one or a few estimates of the true phylogeny, without providing any measure of the variability of the estimate. For this purpose, Felsenstein (1985) introduced the use of the bootstrap (Efron 1982) in phylogenetic inference. Bootstrap samples are created by sampling *sites*, with replacement, from the observed sites. This creates pseudo-sequences of the same length, preserving the selected columns from the original data. For each bootstrap sample, a phylogeny is estimated using a specific phylogenetic construction method. The sampling distribution of the estimated phylogeny is approximated by the empirical distribution of the bootstrap estimates. The method is appropriate under the assumption that the evolutionary histories of the individual sites are independent. Use of this method for phylogenetic inference is increasingly seen in the phylogeny reconstruction literature. The bootstrap technique has been criticized as biased in how bootstrapping estimates the probability distribution of the estimated phylogeny (e.g. Zharkikh and Li 1992a,b, Hillis and Bull 1993). However, Efron, Halloran, and Holmes (1995) showed that the relationship between phylogenies based on the bootstrap samples and the estimated phylogeny mimics the relationship between the estimated and true phylogenies. Thus, the bootstrap does, in fact, correctly estimate the variability of the estimated phylogeny. Of course, as is always the case with the bootstrap method, no information is provided about whether the estimation technique being studied is itself biased.

In the following chapters we introduce a new phylogenetic reconstruction method. The method generates a group of phylogenetic trees using the Markov chain Monte

Carlo technique.

In chapter 2 we first derive a conditional probability distribution of phylogeny given observed sequences of taxa under a variety of evolutionary models. In chapter 3 we propose two Markov chains, each based on the Metropolis-Hastings algorithm. For each chain, we prove that the generated trees converge to the conditional probability distribution of the phylogeny given the observed sequences. In that chapter we also discuss how the Markov chain Monte Carlo method for Jukes-Cantor's one-parameter molecular evolutionary model is implemented in computer programs. In chapter 4 we assess how well the Markov chain Monte Carlo method works. First we assess the Markov chain Monte Carlo method by comparing its estimated probabilities to the true probabilities in the $n = 3$ taxa case. When a direct calculation from the true distribution becomes infeasible for $n \geq 4$, we use two diagnostic tests. The scaled regeneration quantile plot (Mykland, Tierney, and Yu 1995) is used to do a primary check whether on the Markov chain is trapped in a single or a few topologies. We also give another diagnostic to detect divergence of chains. For this diagnostic purpose, we generate two chains starting at two different initial trees. If both chains converge, they should converge to the same stationary distribution. To prove two chains converge to the same distribution, we appeal to the theorem in Maa, Pearl, and Bartoszynski (1996). Based on the theorem we can reduce the comparison between distributions of two multidimensional trees into the comparison of distributions of one-dimensional distances between trees within each chain and between trees from the different chains. Next in this chapter we discuss three estimates of the true phylogeny and the variability of the estimates. In chapter 5 we apply the Markov chain Monte Carlo method to generate phylogenetic trees for the Bacteriophage T7 experimental

data, for data set involving sequences from 16 species of fish, and for a data set involving viral sequences from Russian children with the Human Immunodeficiency Virus (HIV).

In a very recent paper Mau, Newton, and Larget (1996) also propose a phylogenetic reconstruction method based on Markov chain Monte Carlo. Their algorithm is quite different from ours. In Chapter 6 we compare their method with ours and discuss some extensions.

CHAPTER 2

A PROBABILITY DISTRIBUTION OF THE PHYLOGENETIC TREE

2.1 Notation

Each phylogenetic tree consists of three components: a branching pattern (or topology), nodes, and branch lengths. We consider the case where phylogenies are represented by rooted trees and each node is associated with a DNA sequence.

Let $\mathcal{T}_n$ be the set of branching patterns with n external nodes. The n taxa are labeled with $1, 2, \dots, n$. For each phylogenetic tree in $\mathcal{T}_n$ there are $n - 1$ internal nodes, which we label with the negative integers $-1, -2, -3, \dots, -(n - 1)$. Denote the vector of DNA sequences of the external nodes by $\mathbf{b} = (b_1, b_2, \dots, b_n)$ and the DNA sequences of the internal nodes by $\mathbf{v} = (v_{-1}, v_{-2}, \dots, v_{-(n-1)})$. Assume each sequence is m nucleotides long. Finally, let $\mathcal{D}$ be the set of four nucleotides, A, G, C, and T. Thus $b_j \in \mathcal{D}^m$, $j = 1, 2, \dots, n$ and $v_{-i} \in \mathcal{D}^m$, $i = 1, 2, \dots, n - 1$.

Any rooted tree defines not only a root but also the unique path from the root to each external node. The direction of each path corresponds to evolutionary time. Each internal node indicates an evolutionary splitting. Initially we will assume that the time of the whole evolutionary process, τ , from the root to the external nodes,

is a fixed parameter. Let -1 be the root node and $\mathbf{t} = (t_{-2}, t_{-3}, \dots, t_{-(n-1)})$ be the vector of evolutionary splitting times at the internal nodes $-2, -3, \dots, -(n-1)$, respectively. The evolutionary splitting time at each internal node must be less than τ . Thus $\mathbf{t}$ is a vector variable with range $(0, \tau)^{n-2}$. For convenience, we define the times of the external nodes as zero. Branch lengths are defined as the time difference between the two nodes which are connected by the branch.

In summary, a quadruple $(T, \mathbf{t}, \mathbf{v}, \mathbf{b})$, where $T \in \mathcal{T}_n$, $\mathbf{t} \in (0, \tau)^{n-2}$, $\mathbf{v} \in \mathcal{D}^{(n-1)m}$, and $\mathbf{b} \in \mathcal{D}^{nm}$, represents a phylogenetic tree, so that the space of phylogenetic trees with n external nodes is $\Omega_n = \mathcal{T}_n \times \mathcal{D}^{(2n-1)m} \times (0, \tau)^{(n-2)}$. The natural σ -field on Ω_n , $\mathcal{F}_n$, is then the product field of the three Borel fields on $\mathcal{T}_n$, $(0, \tau)^{n-2}$, and $\mathcal{D}^{(2n-1)m}$, respectively.

2.2 A probability distribution of the phylogenetic tree

In order to fix the probability distribution, we make two key assumptions:

Assumption 1 [Neutrality] Substitutions at each site occur according to a homogeneous Poisson process.

Assumption 2 [Independence] Substitutions at different sites are independent.

The probability distribution on the set of trees can be written as the product of three probabilities as follows:

$$P(T, \mathbf{t} \in \Delta, \mathbf{v}, \mathbf{b}) = P(\mathbf{v}, \mathbf{b} \mid (T, \mathbf{t} \in \Delta))P(\mathbf{t} \in \Delta \mid T)P(T) \quad (2.1)$$

Under Assumptions 1 and 2, the unconditional distribution of the topology T , with n external nodes, is the uniform distribution over the $\prod_{k=2}^n (2k-3)$ topologies in

$\mathcal{T}_n$. Suppose that the stationary probabilities of four nucleotides in $\mathcal{D}$ are π_A , π_G , π_C , and π_T , respectively. When the Poisson processes have reached their stationary state, the unconditional distribution of v 's and b 's at each site is the stationary probability of nucleotides in $\mathcal{D}$. However, to fully specify the unconditional distribution of a tree, we need to find the distribution of t given T and the distribution of (v, b) given T and t .

Figure 2.1: Two different patterns in $\mathcal{T}_4$.

When the branching pattern is given, the ancestor-offspring relationships are established and the times of the ancestral nodes must be longer than the times of their offspring. We use this fact to obtain a formula for the probability distribution of the times given the topology, in Lemma 1 below. To motivate the formula, consider the two trees in Figure 2.1. In Figure 2.1(a), the root node, -1 , is an ancestor of internal nodes -2 and -3 , while internal node -2 is an ancestor of internal node -3 alone. Thus the time vector $t = (t_{-2}, t_{-3})$ satisfies $t_{-3} \leq t_{-2} \leq \tau$. Under the two assumptions the splitting time is uniformly distributed over $(0, \tau)$. Thus the

probability density of $\mathbf{t} = (t_{-2}, t_{-3})$ is $\frac{2}{\tau^2}$ over the set $t_{-3} \leq t_{-2} \leq \tau$. On the other hand, in Figure 2.1(b) there is no ancestor-offspring relationship between node -2 and node -3 , although both of them are offspring of node -1 . So t_{-2} and t_{-3} should be independently distributed given the value of τ . Thus

$$f(\mathbf{t} | T) = f(t_{-2}, t_{-3} | T) = f(t_{-2} | T)f(t_{-3} | T) = \frac{1}{\tau^2} \quad \text{for } t_{-2} \leq \tau \text{ and } t_{-3} \leq \tau.$$

Lemma 1 *The probability density of $\mathbf{t} = (t_{-2}, t_{-3}, \dots, t_{-(n-1)})$ given a topology*

$T \in \mathcal{T}_n$, for $n \geq 3$, is

$$f(\mathbf{t} | T) = \frac{1}{\tau^{n-2}} \prod_{i=2}^{n-1} (n_{-i} + 1) \quad \text{for } \mathbf{t} \text{ in the ordered set specified by } T,$$

where n_{-i} , $i = 2, 3, \dots, n-1$, is the number of internal nodes descending from the internal node $-i$, $i = 2, 3, \dots, n-1$.

Proof: The proof of this Lemma uses an induction argument, assuming its validity for the left-branch subtree and the right-branch subtree at the root.

When $n = 3$, there is only one type of tree pattern. So $\mathbf{t} = \{t_{-2}\}$ and its conditional distribution given $T \in \mathcal{T}_3$ is

$$f(\{t_{-2}\} | T) = \frac{1}{\tau} \quad \text{for } 0 \leq t_{-2} \leq \tau.$$

Since $n_{-2} = 0$ and $\frac{1}{\tau^{(n-2)}} \prod_{i=2}^{n-1} (n_{-i} + 1) = \frac{1}{\tau}$, Lemma 1 holds for $n = 3$.

Suppose that it holds for any $k < n$. We will prove that it holds for $k = n$.

For any tree $T \in \mathcal{T}_n$, its root has two children, say s and l . The nodes s and l are either two internal nodes or one internal node and one external node. If they are two internal nodes, each of them leads to a subtree of size less than n . Suppose that T_1

is the subtree descending from the node s of size n_1 and T_2 is the subtree descending from the node l of size n_2 where $n_1 + n_2 = n$. $\mathbf{t}'$ is the splitting times at the internal nodes after t_s in subtree T_1 and $\mathbf{t}''$ is the splitting times at the internal nodes after t_l in subtree T_2 . So $\mathbf{t} = (t_s, \mathbf{t}', t_l, \mathbf{t}'')$. Each coordinate of $\mathbf{t}'$ should be less than t_s and each coordinate of $\mathbf{t}''$ should be less than t_l . $(t_s, \mathbf{t}')$ and $(t_l, \mathbf{t}'')$ are independently distributed. By induction

$$f(\mathbf{t}' | T_1, t_s) = \frac{1}{t_s^{n_1-2}} \prod_{-i \in T_1} (n_{-i} + 1) \quad \text{for } \mathbf{t}' \text{ in the ordered set by } T_1$$

and

$$f(\mathbf{t}'' | T_2, t_l) = \frac{1}{t_l^{n_2-2}} \prod_{-j \in T_2} (n_{-j} + 1) \quad \text{for } \mathbf{t}'' \text{ in the ordered set by } T_2$$

where n_{-i} is the number of internal nodes descending from the internal node $-i$ in T_1 and n_{-j} is the number of internal nodes descending from the internal node $-j$ in T_2 . The values of n_{-i} , $-i \in T_1$ and n_{-j} , $-j \in T_2$ remain the same in T .

Next, t_s given T_1 is distributed like the $(n_1 - 1)^{\text{th}}$ order statistic of $n_1 - 1$ uniformly distributed random variables. Similarly, t_l given T_2 is distributed like the $(n_2 - 1)^{\text{th}}$ order statistic of $n_2 - 1$ uniformly distributed random variables. And t_s and t_l given T are independently distributed. So

$$f(t_s | T) = f(t_s | T_1) = (n_1 - 1) \frac{t_s^{n_1-2}}{\tau^{n_1-1}} \quad \text{for } t_s \in (0, \tau)$$

and

$$f(t_l | T) = f(t_l | T_2) = (n_2 - 1) \frac{t_l^{n_2-2}}{\tau^{n_2-1}} \quad \text{for } t_l \in (0, \tau).$$

Therefore

$$\begin{aligned}
f(\mathbf{t} \mid T) &= f(\mathbf{t}' \mid T, t_s) f(t_s \mid T) f(\mathbf{t}'' \mid T, t_l) f(t_l \mid T) \\
&= f(\mathbf{t}' \mid T_1, t_s) f(t_s \mid T) f(\mathbf{t}'' \mid T_2, t_l) f(t_l \mid T) \\
&= \frac{1}{t_s^{n_1-2}} \prod_{-i \in T_1} (n_{-i} + 1)(n_1 - 1) \frac{t_s^{n_1-2}}{\tau^{n_1-1}} \frac{1}{t_l^{n_2-2}} \prod_{-j \in T_2} (n_{-j} + 1)(n_2 - 1) \frac{t_l^{n_2-2}}{\tau^{n_2-1}} \\
&= \prod_{-i \in T_1} (n_{-i} + 1)(n_1 - 1) \prod_{-j \in T_2} (n_{-j} + 1)(n_2 - 1) \frac{1}{\tau^{n_1+n_2-2}} \\
&= \frac{1}{\tau^{n-2}} \prod_{-i \in T} (n_{-i} + 1)
\end{aligned}$$

The last equation holds since $n_s = n_1 - 2$, $n_l = n_2 - 2$, and $n_1 + n_2 = n$. Thus Lemma 1 holds.

If one of the nodes s and l is an external node, T is decomposed into one subtree of size $n - 1$ and one branch. This is a special case in the first scenario. ■

Given T and $\mathbf{t}$, imagine that nucleotide sequences of the nodes are evolved from the nucleotide sequence of the root along the tree. During this evolutionary process, changes in the nucleotide sequence of the root could happen at any site, therefore giving rise to some new sequences. These changes are called mutations. Mutations are classified by the types of changes into (1) *substitutions*, the replacement of one nucleotide by the another, (2) *deletions*, the removal of one or more nucleotides from the sequence, (3) *insertions*, the addition of one or more nucleotides to the sequence, and (4) *inversion*, the rotation by 180° of a double-stranded DNA segment comprised of two or more base pairs (see page 12 of Li and Graur 1990). We only consider one of the types of the changes, i.e. substitutions, in our model.

Given the assumption of independence of evolution at different sites, the probability of change from one nucleotide sequence into another sequence is a product of probabilities of nucleotide changes over m sites. And the conditional probability of

$(\mathbf{v}, \mathbf{b})$ given $T \in \mathcal{T}_n$ and $\mathbf{t} \in (0, \tau)^{(n-2)}$ is the product of the probabilities of change in each tree branch, times the stationary probabilities $\pi_{v_{-1}}$ of the root sequence v_{-1} .

Putting these three probabilities together, we see that the probability distribution of the entire phylogenetic tree, $(T, \mathbf{t}, \mathbf{v}, \mathbf{b}) \in \Omega_n$, is

$$\begin{aligned} P(T, \mathbf{t} \in \Delta, \mathbf{v}, \mathbf{b}) &= P(\mathbf{v}, \mathbf{b} \mid (T, \mathbf{t} \in \Delta)) P(\mathbf{t} \in \Delta \mid T) P(T) \\ &= \left(\int_{\mathbf{t} \in \Delta} \pi_{v_{-1}} \prod_{(i,j) \in E_T} P_{v_i v_j}(t_i - t_j) f(\mathbf{t} \mid T) d\mathbf{t} \right) \prod_{k=2}^n \frac{1}{(2k-3)}, \end{aligned} \quad (2.2)$$

where E_T is the set of branches (or edges) of T , in which each branch is denoted by the two connected nodes, e.g. (i, j) represents the branch that connects node i and node j , where node i is an ancestor of node j , and $P_{v_i v_j}(t_i - t_j)$ is the probability that the nucleotide sequence at node i changes to the nucleotide sequence at node j over a period of time $t_i - t_j$. ($\pi_{v_{-1}}$ in (2.2) arises from the unconditional distribution of the sequence at the root node.) Note the relaxed notation in expression (2.2), where we have used v 's to represent the nucleotide sequences for all nodes. For the external nodes, the v 's should be thought of as b 's. Also note that the integration of $P(\mathbf{v}, \mathbf{b} \mid T, \mathbf{t})$ in (2.2) over all sequences of internal nodes is the likelihood function in the maximum likelihood method (Felsenstein 1981).

Theorem 1 *The measure $P(\cdot)$ defined in (2.2) is a probability measure on the σ -field $\mathcal{F}_n$.*

Proof:

(i) $P(\cdot) \geq 0$ is obvious.

(ii) Given a $T \in \mathcal{T}_n$,

$$P((T, \mathbf{t}, \mathbf{v}, \mathbf{b}), \mathbf{t} \in (0, \tau)^{n-2}, \mathbf{v} \in \mathcal{D}^{(n-1)m}, \mathbf{b} \in \mathcal{D}^{nm})$$

$$\begin{aligned}
&= \int_{\mathbf{t} \in (0, \tau)^{(n-2)}} \sum_{\mathbf{v} \in \mathcal{D}^{(n-1)m}} \sum_{\mathbf{b} \in \mathcal{D}^{nm}} P(T, \mathbf{t}, \mathbf{v}, \mathbf{b}) d\mathbf{t} \\
&= \int_{\mathbf{t} \in (0, \tau)^{(n-2)}} \sum_{\mathbf{v} \in \mathcal{D}^{(2n-1)m}} \pi_{v_{-1}} \prod_{(i,j) \in E_T} P_{v_i v_j}(t_i - t_j) f(\mathbf{t} | T) \frac{1}{\prod_{k=2}^n (2k-3)} d\mathbf{t} \\
&= \frac{1}{\prod_{k=2}^n (2k-3)} \int_{\mathbf{t} \in (0, \tau)^{(n-2)}} f(\mathbf{t} | T) \sum_{\mathbf{v} \in \mathcal{D}^{(2n-1)m}} \pi_{v_{-1}} \prod_{(i,j) \in E_T} P_{v_i v_j}(t_i - t_j) d\mathbf{t},
\end{aligned} \tag{2.3}$$

We first prove that $\sum_{\mathbf{v} \in \mathcal{D}^{(2n-1)m}} \pi_{v_{-1}} \prod_{(i,j) \in E_T} P_{v_i v_j}(t_i - t_j) = 1$. Since all sites mutate independently, $\sum_{\mathbf{v} \in \mathcal{D}^{(2n-1)m}} \pi_{v_{-1}} \prod_{(i,j) \in E_T} P_{v_i v_j}(t_i - t_j)$ is a product of m summations, which each looks like

$$\sum_{\mathbf{v} \in \mathcal{D}^{(2n-1)}} \pi_{v_{-1}} \prod_{(i,j) \in E_T} P_{v_i v_j}(t_i - t_j) \tag{2.4}$$

We will prove that the summation at each site is equal to one. Then the product of m 1s is one.

Based on the fact that $\sum_{j \in \mathcal{D}} P_{ij}(t) = 1$ for any $i \in \mathcal{D}$ and each external node is connected to only one internal node, if we first sum over four possible states for each external node and fix the states for internal nodes for a moment, then any term related with external nodes disappears from the summation (2.4), i.e.

$$\sum_{\mathbf{v} \in \mathcal{D}^{(2n-1)}} \pi_{v_{-1}} \prod_{(i,j) \in E_T} P_{v_i v_j}(t_i - t_j) = \sum_{\mathbf{v} \in \mathcal{D}^{(n-1)}} \pi_{v_{-1}} \prod_{(i,j) \in E'_T} P_{v_i v_j}(t_i - t_j)$$

where E'_T contains all branches of E_T except the n branches that connect external nodes. In E'_T the nodes that were once connected to external nodes become external nodes in the new tree. Similarly if we do summations for new external nodes as we did for the old external nodes, the terms related with those nodes disappear from the

expression (2.4) and the branches connected to them disappear from E' . We repeat this summation scheme until there are only two branches, say branch $(-1, -2)$ and branch $(-1, -3)$, are left. We summarize all those steps as the follows:

$$\begin{aligned}
& \sum_{\mathbf{v} \in \mathcal{D}^{(2n-1)}} \pi_{v_{-1}} \prod_{(i,j) \in E_T} P_{v_i v_j}(t_i - t_j) = \sum_{\mathbf{v} \in \mathcal{D}^{(n-1)}} \pi_{v_{-1}} \prod_{(i,j) \in E'_T} P_{v_i v_j}(t_i - t_j) \\
& = \dots \\
& = \sum_{v_{-1}, v_{-2}, v_{-3} \in \mathcal{D}} \pi_{v_{-1}} P_{v_{-1} v_{-2}}(\tau - t_{-2}) P_{v_{-1} v_{-3}}(\tau - t_{-3}) \\
& = \sum_{v_{-1} \in \mathcal{D}} \pi_{v_{-1}} \left(\sum_{v_{-2} \in \mathcal{D}} P_{v_{-1} v_{-2}}(\tau - t_{-2}) \right) \left(\sum_{v_{-3} \in \mathcal{D}} P_{v_{-1} v_{-3}}(\tau - t_{-3}) \right) \\
& = \sum_{v_{-1} \in \mathcal{D}} \pi_{v_{-1}} \\
& = 1
\end{aligned} \tag{2.5}$$

Substituting (2.5) into (2.3), we have

$$\begin{aligned}
& P((T, \mathbf{t}, \mathbf{v}, \mathbf{b}) \text{ where } \mathbf{t} \in (0, \tau)^{n-2}, \mathbf{v} \in \mathcal{D}^{(n-1)m}, \mathbf{b} \in \mathcal{D}^{nm}) \\
& = \int_{\mathbf{t}} \frac{1}{\prod_{k=1}^{n-1} (2k-3)} f(\mathbf{t}) d\mathbf{t} = \frac{1}{\prod_{k=1}^{n-1} (2k-3)}.
\end{aligned} \tag{2.6}$$

Hence

$$\begin{aligned}
P(\Omega_n) &= P((T, \mathbf{t}, \mathbf{v}, \mathbf{b}) \text{ where } \mathbf{t} \in (0, \tau)^{n-2}, \mathbf{v} \in \mathcal{D}^{(n-1)m}, \mathbf{b} \in \mathcal{D}^{nm}, T \in \mathcal{T}_n) \\
&= \sum_{T \in \mathcal{T}_n} \frac{1}{\prod_{k=1}^{n-1} (2k-3)} = 1.
\end{aligned}$$

(iii) (Additivity) Proof of the additivity of P is straightforward. ■

Notice that in Theorem 1, we do not explicitly specify $P_{v_i v_j}(t_i - t_j)$, i.e. the probability of substitution from sequence i to sequence j . The theorem is applicable for any substitution model. Numerous statistical models have been proposed in the molecular

evolution literature. We will review four of them: Jukes and Cantor's (1969) *one-parameter model*, Kimura's (1980) *two-parameter model*, Felsenstein's (1981) *four-parameter model*, and Hasegawa, Kishino and Yano's (1985) *five-parameter model*. Each of these models are special cases of the general model studied by Rodriguez et al. (1990). The model allows different substitution rates for each combination of nucleotides. Therefore there are twelve parameters in the model. The 12-parameter model is generally considered an overfit to most molecular data. The five-parameter model, which allows for differing rates of substitution for transitions and transversions as well as differing equilibrium probabilities for each nucleotide, is a good compromise.

Jukes and Cantor's one-parameter model

This is the simplest model. It assumes the same substitution rate among the four nucleotides (shown in Figure 2.2 adapted from Li and Graur 1991).

Figure 2.2: One-parameter model of nucleotide substitution. Substitution rate in each direct is α .

A consequence of this assumption is that the probability that nucleotide d_1 at a specific site mutates to nucleotide d_2 over a time period t is

$$\begin{cases} \frac{1}{4}(1 + 3e^{-4\alpha t}) & \text{if } d_1 = d_2 \\ \frac{1}{4}(1 - e^{-4\alpha t}) & \text{if } d_1 \neq d_2 \end{cases} \quad (2.7)$$

for $d_1, d_2 \in \mathcal{D}$. (See for example Chapter 3 in Li and Graur 1991.) Thus, the probability that the nucleotide sequence at node i changes to the nucleotide sequence at node j over a period of time t is

$$P_{v_i v_j}(t) = \left(\frac{1}{4}(1 + 3e^{-4\alpha t})\right)^{m-m_{ij}} \left(\frac{1}{4}(1 - e^{-4\alpha t})\right)^{m_{ij}},$$

where m_{ij} is the number of sites that differ between v_i and v_j .

Kimura's two-parameter model

The assumption that all nucleotide substitutions occur equally likely, as in Jukes and Cantor's one parameter model, is unrealistic in most cases. For example, transitions between A and G or between C and T are generally more frequent than transversions between one Purine and one Pyrimidine. To take this fact into account, Kimura (1980) proposed a two-parameter model which assumes the transition rate ($A \leftrightarrow G$ and $C \leftrightarrow T$) is α and the transversion rate ($A \leftrightarrow C, A \leftrightarrow T, G \leftrightarrow C$, and $G \leftrightarrow T$) is β . A consequence of the assumption is that the probability that the nucleotide d_1 at a specific site mutates to d_2 over a time period t is

$$\begin{cases} \frac{1}{4}(1 + e^{-4\beta t} + 2e^{-2(\alpha+\beta)t}) & \text{if } d_1 = d_2 \\ \frac{1}{4}(1 + e^{-4\beta t} - 2e^{-2(\alpha+\beta)t}) & \text{if a transition} \\ \frac{1}{4}(1 - e^{-4\beta t}) & \text{if a transversion} \end{cases} \quad (2.8)$$

(See for example Chapter 3 in Li and Graur 1991 or Kimura 1980). Thus, the probability that the nucleotide sequence at node i changes to the nucleotide sequence at

node j over a period of time t is

$$P_{v_i v_j}(t) = \left(\frac{1}{4}\right)^m (1 + e^{-4\beta t} + 2e^{-2(\alpha+\beta)t})^{m-m_{ij}-n_{ij}} (1 + e^{-4\beta t} - 2e^{-2(\alpha+\beta)t})^{m_{ij}} (1 - e^{-4\beta t})^{n_{ij}}$$

where m_{ij} is the number of transitions between v_i and v_j and n_{ij} is the number of transversions between v_i and v_j .

Felsenstein's four-parameter model

Felsenstein's four-parameter model assumes that the current base at a site is replaced at a constant rate θ . The replacement is A, G, C, or T with probabilities π_A , π_G , π_C , or π_T . A consequence of this assumption is that the probability that the nucleotide d_1 at a specific site mutates to d_2 over a time period t is

$$e^{-\theta t} \delta_{d_1 d_2} + (1 - e^{-\theta t}) \pi_{d_2} \quad (2.9)$$

where

$$\delta_{ij} = \begin{cases} 1 & \text{if } i = j \\ 0 & \text{if } i \neq j \end{cases}$$

(see Felsenstein 1981). Thus, the probability that the nucleotide sequence at node i changes to the nucleotide sequence at node j over a period of time t is

$$P_{v_i v_j}(t) = \prod_{b \in \mathcal{D}} \left(e^{-\theta t} + (1 - e^{-\theta t}) \pi_b \right)^{m_{bb}} \prod_{b \in \mathcal{D}} \left((1 - e^{-\theta t}) \pi_b \right)^{m_b - m_{bb}}$$

where m_{bb} , $b \in \mathcal{D}$, is the number of sites where $v_i = v_j = b$, $b \in \mathcal{D}$ and m_b , $b \in \mathcal{D}$, is the number of sites where $v_j = b$, $b \in \mathcal{D}$.

Hasegawa, Kishino, and Yano's five-parameter model

Hasegawa, Kishino, and Yano's five-parameter model is an extension of Felsenstein's four-parameter model. The model further allows for two different replacement

rates. For example, the replacement rate between two purines or two pyrimidines (transition) differs from the replacement rate between one Purine and one Pyrimidine (transversion). It assumes the ratio of the two replacements rates is κ . A consequence of this assumption is that the transition probability matrix among the four nucleotides (A, G, C, T) is

$$\begin{aligned}
& \begin{pmatrix} \frac{\pi_A}{\pi_R}e^{-\theta t} + \frac{\pi_G}{\pi_R}e^{-\gamma_1 t} & \frac{\pi_G}{\pi_R}(e^{-\theta t} - e^{-\gamma_1 t}) & 0 & 0 \\ \frac{\pi_A}{\pi_R}(e^{-\theta t} - e^{-\gamma_1 t}) & \frac{\pi_G}{\pi_R}e^{-\theta t} + \frac{\pi_A}{\pi_R}e^{-\gamma_1 t} & 0 & 0 \\ 0 & 0 & \frac{\pi_C}{\pi_Y}e^{-\theta t} + \frac{\pi_T}{\pi_Y}e^{-\gamma_2 t} & \frac{\pi_T}{\pi_Y}(e^{-\theta t} - e^{-\gamma_2 t}) \\ 0 & 0 & \frac{\pi_C}{\pi_Y}(e^{-\theta t} - e^{-\gamma_2 t}) & \frac{\pi_T}{\pi_Y}e^{-\theta t} + \frac{\pi_C}{\pi_Y}e^{-\gamma_2 t} \end{pmatrix} \\
& + (1 - e^{-\theta t}) \begin{pmatrix} \pi_A & \pi_G & \pi_C & \pi_T \\ \pi_A & \pi_G & \pi_C & \pi_T \\ \pi_A & \pi_G & \pi_C & \pi_T \\ \pi_A & \pi_G & \pi_C & \pi_T \end{pmatrix}
\end{aligned} \tag{2.10}$$

where $\pi_R = \pi_A + \pi_G$, $\pi_Y = \pi_C + \pi_T$, $\gamma_1 = \theta(\kappa\pi_R + \pi_Y)$, and $\gamma_2 = \theta(\kappa\pi_Y + \pi_R)$ (see for example Hasegawa, Kishino, and Yano 1985 and Mau, Newton, and Larget 1996). Thus, the probability that the nucleotide sequence at node i changes to the nucleotide sequence at node j over a period of time t is

$$\begin{aligned}
P_{v_i v_j}(t) = & \left(\frac{\pi_A \pi_Y}{\pi_R} e^{-\theta t} + \frac{\pi_G}{\pi_R} e^{-\gamma_1 t} + \pi_A \right)^{m_{AA}} \left(\frac{\pi_G \pi_Y}{\pi_R} e^{-\theta t} + \frac{\pi_A}{\pi_R} e^{-\gamma_1 t} + \pi_G \right)^{m_{GG}} \\
& \left(\frac{\pi_C \pi_R}{\pi_Y} e^{-\theta t} + \frac{\pi_T}{\pi_Y} e^{-\gamma_2 t} + \pi_C \right)^{m_{CC}} \left(\frac{\pi_T \pi_R}{\pi_Y} e^{-\theta t} + \frac{\pi_C}{\pi_Y} e^{-\gamma_2 t} + \pi_T \right)^{m_{TT}} \\
& \left(\frac{\pi_G}{\pi_R} (e^{-\theta t} - e^{-\gamma_1 t}) \right)^{m_{AG}} \left(\frac{\pi_A}{\pi_R} (e^{-\theta t} - e^{-\gamma_1 t}) \right)^{m_{GA}} \left(\frac{\pi_C}{\pi_Y} (e^{-\theta t} - e^{-\gamma_2 t}) \right)^{m_{CT}} \\
& \left(\frac{\pi_T}{\pi_Y} (e^{-\theta t} - e^{-\gamma_2 t}) \right)^{m_{TC}} \prod_{b \in \mathcal{D}} \left((1 - e^{-\theta t}) \pi_b \right)^{n_b}
\end{aligned}$$

where m_{sl} , $s, l \in \mathcal{D}$, is the number of sites that $v_i = s$ and $v_j = l$ and n_b , $b \in \mathcal{D}$, is the number of transversions if the current base is b .

The first three models are three special cases of the five-parameter model. The one-parameter model is the case when $\pi_A = \pi_G = \pi_C = \pi_T = \frac{1}{4}$, $\kappa = 1$, and $\theta = 4\alpha$. The two-parameter model is the case when $\pi_A = \pi_G = \pi_C = \pi_T = \frac{1}{4}$, $\kappa = \frac{\alpha}{\beta}$, and $\theta = 4\beta$. While the four-parameter model is the case when $\kappa = 1$.

2.3 Conditional distribution of the phylogenetic tree given sequences of taxa

In phylogenetic studies the nucleotide sequences of the external nodes are observable. So the sample space is restricted to the subset of the tree space in which $\mathbf{b}$ matches the observed sequences of external nodes. For instance if we want information about the tree topologies, the appropriate conditional distribution is

$$P(T | \mathbf{b}) = \frac{P(T, \mathbf{b})}{P(\mathbf{b})} = \frac{\int_{\mathbf{t} \in (0, \tau)^{(n-2)}} \sum_{\mathbf{v} \in \mathcal{D}^{mn}} P(T, \mathbf{t}, \mathbf{v}, \mathbf{b}) d\mathbf{t}}{\sum_{T \in \mathcal{T}_n} \int_{\mathbf{t} \in (0, \tau)^{(n-2)}} \sum_{\mathbf{v} \in \mathcal{D}^{mn}} P(T, \mathbf{t}, \mathbf{v}, \mathbf{b}) d\mathbf{t}}. \quad (2.11)$$

The denominator in (2.11) is a sum of $\prod_{k=2}^n (2k-3)$ high-dimensional integrals. The number of such integrals grows extremely fast. For example, when $n = 10$, it is 34,459,425. Thus the direct calculation of the conditional probability distribution of a topology given the observed nucleotide sequences is essentially infeasible and inefficient, because calculation of each integral is already computationally demanding. However the denominator in (2.11) is only a normalizing constant. For any distribution ν , known only up to a normalizing constant, the Metropolis-Hastings algorithm

(see, e.g. Tierney 1994) enables the construction of a class of Markov chains with stationary distribution ν .

CHAPTER 3

METROPOLIS ALGORITHM

In this chapter we construct two Markov chains whose stationary distributions are the conditional distribution of the phylogenetic tree given the observed sequences. We show that the first chain is “uniformly ergodic” and the second chain is “ergodic”, so that samples generated from them enable us to infer stochastic properties of the phylogenetic history.

3.1 Metropolis-Hastings sampler review

In this section we give a brief review of the construction of the transition probability function $P(\cdot, \cdot)$ of the Metropolis-Hastings sampler and convergence theorems. Let Π be a probability distribution on a measurable space $(\mathcal{X}, \mathcal{B})$. Assume that Π has a density π with respect to a measure μ . Let P be a transition probability function on a measurable space $(\mathcal{X}, \mathcal{B})$, i.e. P is a function on $\mathcal{X} \times \mathcal{B}$ such that for each $x \in \mathcal{X}$, $P(x, \cdot)$ is a probability measure on $(\mathcal{X}, \mathcal{B})$, and for each $C \in \mathcal{B}$, $P(\cdot, C)$ is a measurable function on $(\mathcal{X}, \mathcal{B})$. Suppose that Π is a stationary probability of P . A Markov chain $\{X_n; n \geq 1\}$ is generated from P . If the transition probability P satisfies certain conditions, the Markov chain generated from P is ergodic, i.e. the distribution of X_n converges to π .

Let $T(C) = \inf\{n : n > 0, X_n \in C\}$ be the first time the chain hits C . The following definitions and the results apply to any Markov chain Monte Carlo algorithm.

Definitions: (i) A probability measure, π , is *invariant* for the Markov chain generated from a transition probability, P , if

$$\pi(C) = \int P(x, C)\pi(dx) \quad \text{for all } C \in \mathcal{B}.$$

(ii) A set $A \in \mathcal{B}$ is said to be *accessible* from x if $P_x(T(A) < \infty) > 0$. A transition probability P is π -*irreducible* if every $A \in \mathcal{B}$, with $\pi(A) > 0$, is accessible from every $x \in \mathcal{X}$.

(iii) A π -irreducible transition probability P is *aperiodic* if

$$g.c.d.\{m : \text{there is an } \epsilon_m > 0 \text{ such that } P^m(x, \cdot) \geq \epsilon_m \pi(\cdot) \text{ for each } x \in \mathcal{X}\} = 1.$$

(iv) A transition probability P satisfies the *Doebelin condition* if there exists a probability measure ϕ on $(\mathcal{X}, \mathcal{B})$, an integer k , and an $\epsilon > 0$ such that

$$P^k(x, C) \geq \epsilon \phi(C) \quad \text{for all } x \in \mathcal{X} \text{ and all } C \in \mathcal{B}. \quad (3.1)$$

Note: The Doeblin condition is also called the minorization condition on the space $\mathcal{X}$ in Tierney (1994).

Proposition 1. Suppose that the Markov chain $\{X_n; n \geq 1\}$ with transition function $P(x, C)$ has an invariant probability measure π . Assume that the chain is π -irreducible and aperiodic. Then

$$\sup_{C \in \mathcal{B}} |P^N(x, C) - \pi(C)| \rightarrow 0 \quad \text{for } \pi\text{-almost all } x. \quad (3.2)$$

The proof of Proposition 1 can be found in Athreya, Doss, and Sethuraman (1996) or Tierney (1994).

Doob (1953) gave sufficient conditions for a chain to be uniformly ergodic. The chain is uniformly ergodic if there is a positive constant M and a positive constant $0 < r < 1$ such that

$$\sup_{x \in \mathcal{X}} |P^N(x, \cdot) - \pi| \leq Mr^N.$$

The following proposition is equivalent to Doob's theorem.

Proposition 2. Suppose that the Markov chain satisfies the Doeblin condition. Then there exists a unique invariant probability measure π such that for all N

$$\sup_{C \in \mathcal{B}} |P^N(x, C) - \pi(C)| \leq (1 - \epsilon)^{(N/k)-1} \quad \text{for each } x \in \mathcal{X}. \quad (3.3)$$

A short proof of Proposition 2 is reviewed in Doss, Huffer, and Lawson (1994).

A Metropolis-Hastings transition probability P is constructed as follows. Let $Q(\cdot, \cdot)$ be an arbitrary transition probability function on a measure space $(\mathcal{X}, \mathcal{B})$ and let $q(\cdot, \cdot)$ be the density. Let p be the density function of P .

Let

$$p(x, y) = \begin{cases} q(x, y)\alpha(x, y) & \text{if } y \neq x \\ 0 & \text{if } y = x \end{cases}$$

and

$$P(x, dy) = p(x, y)\pi(dy) + r(x)\delta_x(dy)$$

where

$$\alpha(x, y) = \begin{cases} \min\{\frac{\pi(y)q(y, x)}{\pi(x)q(x, y)}, 1\} & \text{if } \pi(x)q(x, y) > 0 \\ 1 & \text{if } \pi(x)q(x, y) = 0 \end{cases} \quad (3.4)$$

and

$$r(x) = 1 - \int p(x, y)\pi(dy).$$

A Metropolis-Hastings transitional probability P satisfies the reversibility condition

$$\pi(x)p(x, y) = \pi(y)p(y, x).$$

It is easy to check that when P satisfies the reversibility condition, π is invariant for the Markov chain generated from P . For P to be irreducible, it is necessary that Q be irreducible, but this is not sufficient: irreducibility of P depends on both Q and π . If P is irreducible and $\pi(\{x : r(x) > 0\}) > 0$, then, by the results of Nummelin (1984), Section 2.4, P is aperiodic. Therefore, by the result of Proposition 1, (3.2) holds.

3.2 Two Metropolis-Hastings samplers on the phylogenetic tree space

In this section we show how to estimate the conditional distribution of topologies using the Metropolis-Hastings algorithm. To avoid the calculation of the integral in the numerator of (2.11), we will sample the whole tree instead of sampling only the topology. We use the fact that generated topologies converge to the appropriate topology distribution, if the generated trees converge to the appropriate tree distribution. The conditional distribution of the topology given the observed nucleotide sequences of external nodes is the appropriate marginal distribution of the tree given the observed nucleotide sequences of external nodes.

In our situation, we take π to be the conditional density of the tree given the observed nucleotide sequences of external nodes, i.e.

$$\pi(T, \mathbf{t}, \mathbf{v}) = p(T, \mathbf{t}, \mathbf{v} \mid \mathbf{b}) = \frac{p(T, \mathbf{t}, \mathbf{v}, \mathbf{b})}{\sum_{T \in \mathcal{T}_n} \int_{\mathbf{t} \in (0, \tau)^{(n-2)}} \sum_{\mathbf{v} \in \mathcal{D}^{mn}} p(T, \mathbf{t}, \mathbf{v}, \mathbf{b}) d\mathbf{t}}, \quad (3.5)$$

where p is the density function associated with P of (2.2).

A minimal requirement for the Markov chain simulation method to work is that the chain be irreducible. We now describe two Metropolis-Hastings algorithms that have this property.

Notice that any two topologies in $\mathcal{T}_n$ can be transformed from one to another by small perturbations. There are many kinds of perturbations that are possible. We have studied two kinds which make one tree accessible from any other. The first method of transformation moves a “pendant” branch to one of $2n - 3$ possible branches, according to an appropriate probability distribution. We will give the transition probability associated with it and also show that the Markov chain generated by this transition probability converges to the target distribution at a rate that is geometric, uniformly in the starting point.

The second method is essentially the local rearrangement strategy that was used by Kuhner, Yamato, and Felsenstein (1995) in a genealogy application. This rearrangement occurs only in the neighborhood of the internal nodes (except for the fixed root of the tree). In Kuhner, Yamato and Felsenstein (1995), they explained that

“In practice, making larger rearrangements would probably make the sampling less efficient, because if a genealogy already has fairly high probability, a large rearrangement of it is liable to be much worse and therefore be rejected. However, such techniques may prove useful in analyzing very large numbers of sequences, where the chance that the process will become trapped in a local maximum of the posterior probability distribution is greater.”

This is also true in the phylogenetic tree application. The neighborhood of an internal node consists of its parent, brother, and two children. The local rearrangement involves randomly reassigning the three children among parent and internal node. The Markov chain generated by applying this branch rearrangement will also converge to the target distribution.

3.2.1 The pendant branch transition strategy

In Chapter 2 the probability distribution of trees is defined on Ω_n , the space of all rooted trees. Actually the definition is also appropriate for the space of all rooted trees with a prespecified outgroup taxon. In this section we assume that each tree has an outgroup labeled with 0. The number of distinct topologies with n in-group taxa and one outgroup taxon is the same as the number of distinct topologies with n taxa. Without causing any confusion, we denote both sample spaces, a space of trees with n in-group taxa and one outgroup taxon and a space of trees with n taxa, by Ω_n .

In the pendant branch strategy we randomly choose a pendant branch, which connects an external node to its immediate ancestor, remove it from the tree and re-attach it back to a branch such that a valid tree is produced. Given a topology in $\mathcal{T}_n$, there are n pendant branches and there are $2n - 3$ potential branches to which a pendant branch can be re-attached. Consider for example the tree in Figure 3.1(a), which is a topology in $\mathcal{T}_4$. Suppose that the branch $(-2, 2)$ is randomly chosen for a rearrangement. We remove the branch from the tree. The dotted lines in Figure 3.1(b) show all of the possible ways that attachments still make potential trees. But the sixth dotted line violates the assumption that node 0 is the outgroup. Therefore, for

each chosen branch there are only five possible rearrangements. Next we will show that starting with any topology in $\mathcal{T}_n$ we can transform it to any other topology in $\mathcal{T}_n$ by applying at most $n - 2$ such re-attachments.

Figure 3.1: The pendant branch transition strategy. For a tree in $\mathcal{T}_4$, the dotted lines 1–5 show valid possible topologies.

Let the set of topologies which can be transformed from topology T through one such pendant branch rearrangement be $\mathcal{N}_T^1$ and denote the set of topologies which can be transformed from T through some k ($k \leq m$) such rearrangements by $\mathcal{N}_T^m$. It is obvious that $\mathcal{N}_T^1$ is reversible, in the sense that $T \in \mathcal{N}_{T'}^1$, provided that $T' \in \mathcal{N}_T^1$. We have the following result.

Lemma 2 $\mathcal{T}_n \subset \mathcal{N}_T^{n-2}$ for any $T \in \mathcal{T}_n$.

Proof To prove the lemma, it is sufficient to prove that given any pair of topologies in $\mathcal{T}_n$, the number of rearrangements required to change the first topology to the second is at most $n - 2$. This is shown recursively. We begin with only three external nodes, say 0, 1, 2, for which any pair of topologies in $\mathcal{T}_n$ have the same branching pattern. Then imagine that any topology is a result of adding the branches that connect nodes 3, 4, $\dots$, n to the tree. Given a pair of topologies, consider the step of adding a particular branch to both topologies. If this addition gives two different topologies, we rearrange the branch in the first topology so that it matches in the second topology. Such rearrangements will be required for at most $n - 2$ of the steps.

■

Remark Lemma 2 is the key result needed to establish convergence of the pendant branch swapping chain.

Since our Markov chain must generate whole trees from the target distribution π , the Markov chain must take values in Ω_n . Suppose $x = (T_n, \mathbf{t}, \mathbf{v}, \mathbf{b}) \in \Omega_n$. The vector of external nodes, $\mathbf{b}$, remains fixed at the observed values throughout the execution of the algorithm. In order to define the transition probability function $Q_1(x, y)$, we will decompose the pendant branch method of transition into the following three steps.

Step 1 Randomly choose a branch from among the n pendant branches for a rearrangement. Without loss of generality we can call the branch $(-i, i)$. Remove $(-i, i)$ from x and denote the rest of the tree by x^* . x^* is a tree in Ω_{n-1} and it is also a subtree common to all trees which could arise from the re-attachment of the branch. Thus it provides a common factor which can be divided out of $q_1(x, \cdot)$, the density of $Q_1(x, y)$. Next we draw a topology from that subset of $\mathcal{N}_{T_n}^1$ that involves re-attachments of $(-i, i)$. Denote the topology and time vector of x^* by T_{n-1} and $\mathbf{t}^*$,

respectively. Conditional on the selection of the pendant branch $(-i, i)$ we re-attach it to (j, k) , creating topology T'_n . We do this with probability

$$C_{x^*} \int_{t_k}^{t_j} \frac{\sum_{v \in \mathcal{D}^m} P_{vjv}(t_j - t) P_{vv_k}(t - t_k) P_{vb_i}(t)}{P_{vjv_k}(t_j - t_k)} \frac{f(\mathbf{u} \mid T'_n)}{f(\mathbf{t}^* \mid T_{n-1})} dt, \quad (3.6)$$

where $\mathbf{u}$ has t in the i^{th} coordinate and the coordinates of $\mathbf{t}^*$ in the remaining positions, and where C_{x^*} depends on x only through x^* (which is determined by the choice of $(-i, i)$).

Step 2 In this step we decide on the value of t'_{-i} , the time of the attaching point. Given that $(-i, i)$ will be attached on (j, k) in T'_n , we draw t'_{-i} from the probability density

$$\frac{\sum_{v \in \mathcal{D}^m} P_{vjv}(t_j - t'_{-i}) P_{vv_k}(t'_{-i} - t_k) P_{vb_i}(t'_{-i})}{\int_{t_k}^{t_j} \sum_{v \in \mathcal{D}^m} P_{vjv}(t_j - t) P_{vv_k}(t - t_k) P_{vb_i}(t) dt} \quad \text{for } t_k \leq t'_{-i} \leq t_j. \quad (3.7)$$

Step 3 In the last step we decide on a nucleotide sequence for the attaching point. Here, a sequence v'_{-i} is chosen with probability

$$\frac{P_{vjv'_{-i}}(t_j - t'_{-i}) P_{v'_{-i}v_k}(t'_{-i} - t_k) P_{v'_{-i}b_i}(t'_{-i})}{\sum_{v \in \mathcal{D}^m} P_{vjv}(t_j - t'_{-i}) P_{vv_k}(t'_{-i} - t_k) P_{vb_i}(t'_{-i})}. \quad (3.8)$$

From Lemma 1 we notice that the term $f(\mathbf{u} \mid T'_n)$ in (3.6) does not depend on t , allowing it to be factored out of the integral.

For every topology T and sequence $\mathbf{v}$, let $\mu_{T,\mathbf{v}}$ be the probability measure on t associated with the density given in Lemma 1. Let

$$\mu(T_0, \Delta, \mathbf{v}_0) = \sum_{T, \mathbf{v}} \mu_{T,\mathbf{v}}(\Delta) I(T_0 = T, \mathbf{v}_0 = \mathbf{v}). \quad (3.9)$$

The measure μ will serve as the carrier measure for $Q_1(x, y)$.

Denote the resulting tree, after the three steps above, by y . Then the transition probability from x to y has density (with respect to μ) given by the product of (3.6), (3.7) and (3.8), i.e.

$$q_1(x, y) = C_x \cdot \frac{P_{v_j v'_{-i}}(t_j - t'_{-i}) P_{v'_{-i} v_k}(t'_{-i} - t_k) P_{v'_{-i} b_i}(t'_{-i})}{P_{v_j v_k}(t_j - t_k)} \frac{f(\mathbf{t}' | T'_n)}{f(\mathbf{t}^* | T_{n-1})}. \quad (3.10)$$

Conversely if y is an initial state, there is only one way to transform y to x using only one transition, namely moving branch $(-i, i)$ back to the original place. Supposing that it was on (l, s) , then the transition probability from y to x is

$$q_1(y, x) = C_x \cdot \frac{P_{v_l v_{-i}}(t_l - t_{-i}) P_{v_{-i} v_s}(t_{-i} - t_s) P_{v_{-i} b_i}(t_{-i})}{P_{v_l v_s}(t_l - t_s)} \frac{f(\mathbf{t} | T_n)}{f(\mathbf{t}^* | T_{n-1})}. \quad (3.11)$$

Using (2.2), it is easy to prove that $Q_1(x, y)$ satisfies the reversibility condition, i.e. $\pi(x)q_1(x, y) = \pi(y)q_1(y, x)$. Thus the acceptance probability is always one (see (3.4)). Denoting by $P_1(x, y)$ the Metropolis-Hastings transition probability induced by $Q_1(x, y)$, we have $P_1(x, y) = Q_1(x, y)$. The proof of reversibility relies on the fact that the C_x -'s appearing in (3.10) and (3.11) are equal. This is because to go from y to x , we must first choose branch $(-i, i)$ for a rearrangement (which occurs with probability $1/n$) and then re-attach it. After deletion of $(-i, i)$ the resulting tree x^* is identical to the tree x^* that was formed in going from x to y .

(It turns out that if the chain is currently at x , then conditional on having selected $(-i, i)$ for a rearrangement, the distribution of the proposal tree y is the restriction of π to the set of trees that could result from a re-attachment of $(-i, i)$, renormalized to be a probability measure. It is this close relation between q_1 and π that makes this an effective Markov chain.)

In P_1 we often need to calculate the ratio of two distribution densities of time (e.g. in (3.10) and (3.11)). Actually this ratio depends only on the numbers of internal nodes descending from those internal nodes that are ancestors of the attaching point. The ratio can be calculated in short way by using the formula given in Lemma 3.

Lemma 3 *Suppose that x and x^* are as defined above. Then*

$$\frac{f(t \mid T_n)}{f(t^* \mid T_{n-1})} = \frac{1}{\tau} \frac{\prod_{-i \in A_s} (n_{-i} + 2)(n_s + 2)}{\prod_{-i \in A_s} (n_{-i} + 1)},$$

where A_s is the set of ancestral nodes of s and n_{-i} is the number of internal nodes descending from the internal node $-i$ in T_{n-1} . If node s is an external node, then define $n_s = -1$.

Proof.

Based on Lemma 1 we know that $f(t \mid T_n)$ and $f(t^* \mid T_{n-1})$ depends only on the number of internal nodes descending from internal nodes in T_n and in T_{n-1} , respectively. When the removed branch is attached back to T_{n-1} , the numbers of internal nodes descending from the internal nodes that are ancestors of the attaching point are increased by one in the resulting topology T_n . The numbers of internal nodes descending from the rest of the internal nodes remain the same in both T_n and T_{n-1} ; therefore they are canceled from the denominator and the numerator of the ratio. Thus we have

$$\frac{f(t \mid T_n)}{f(t^* \mid T_{n-1})} = \frac{1}{\tau} \frac{\prod_{-i \in A_s} (n_{-i} + 2)(n_s + 2)}{\prod_{-i \in A_s} (n_{-i} + 1)},$$

where $n_s + 2$ in the numerator is the number of internal nodes descending from the attaching node plus one. ■

3.2.2 The local rearrangement transition strategy

Our second method for producing a Markov chain is the local rearrangement transition strategy. This method does not demand that any taxon be declared an outgroup. However, the algorithm can also be used if a prespecified outgroup is known. In particular, when the parent of the randomly chosen target is the root of the tree, there will be two choices out of three which violate the outgroup assumption. These choices are automatically rejected. If the parent node is not the root, the algorithm remains the same.

The local rearrangement transition strategy is also divided into three steps. In the first step, we randomly choose a node from among all nodes that have both a parent and children (i.e. are neither external nodes nor the root of the tree). The chosen node is referred to as the “target.” Next, we pick two sequences from amongst the three children (the target’s two children and the brother) to become the new children of the target (see Figure 3.2). The probability associated with the first step is a constant $\frac{1}{3(n-2)}$ where $\frac{1}{n-2}$ is the probability that one internal node is randomly chosen among the $n-2$ internal nodes and $\frac{1}{3}$ is the probability that a particular pattern in Figure 3.2 (adapted from Kuhner, Yamato, and Felsenstein 1995) is randomly chosen.

If the rearrangement is applied repeatedly, within a finite number of times, any topology can be transformed into any other topology (see Kuhner, Yamato, and Felsenstein 1995). This is the key property of the local rearrangement strategy needed to establish convergence of the Markov chain.

Suppose that x is an initial tree and that y is a tree produced after a local rearrangement. The difference between x and y is in the target’s neighborhood. Let C_1 , C_2 and C_3 be the two children and the brother of the target node, T , in x , and let

Figure 3.2: The local rearrangement transition strategy. A neighborhood is selected for local rearrangement (a), possibly resulting in trees (b) or (c).

C'_1 , C'_2 and C'_3 be the two children and the brother of the target node, T' , in y . In the second step, we decide on the time of the target node in tree y . The time of the target node must be larger than the times of its two children, $t_{C'_1}$ and $t_{C'_2}$, and less than the time of its parent, t_P . The value of $t_{T'}$ is drawn according to the probability distribution of $t_{T'}$ given the sequences and the splitting times of the parent and the two new children for which the density function is

$$\frac{\sum_{v \in \mathcal{D}^m} P_{v_P v_{T'}}(t_P - t_{T'}) P_{v_{T'} v_{C'_1}}(t_{T'} - t_{C'_1}) P_{v_{T'} v_{C'_2}}(t_{T'} - t_{C'_2})}{\int_{\max\{t_{C'_1}, t_{C'_2}\}}^{t_P} \sum_{v \in \mathcal{D}^m} P_{v_P v}(t_P - t) P_{v v_{C'_1}}(t - t_{C'_1}) P_{v v_{C'_2}}(t - t_{C'_2}) dt}.$$

Finally in the last step, we decide on the nucleotide sequence for the target node in tree y given the splitting times and the sequences of the rest of nodes. The nucleotide sequence of the target depends on the sequences of the rest of nodes through the sequences of its parent and two children. The probability distribution of $v_{T'}$ is

$$\frac{P_{v_P v_{T'}}(t_P - t_{T'}) P_{v_{T'} v_{C'_1}}(t_{T'} - t_{C'_1}) P_{v_{T'} v_{C'_2}}(t_{T'} - t_{C'_2})}{\sum_{v \in \mathcal{D}^m} P_{v_P v}(t_P - t) P_{v v_{C'_1}}(t - t_{C'_1}) P_{v v_{C'_2}}(t - t_{C'_2}) dt}.$$

The transition probability from x to y is the product of the probabilities in the three steps above, i.e.

$$q_2(x, y) = \frac{1}{(n-2)} \frac{1}{3} \frac{P_{v_P v_{T'}}(t_P - t_{T'}) P_{v_{T'} v_{C1'}}(t_{T'} - t_{C1'}) P_{v_{T'} v_{C2'}}(t_{T'} - t_{C2'})}{\int_{\max\{t_{C1'}, t_{C2'}\}}^{t_P} \sum_{v \in \mathcal{D}^m} P_{v_P v}(t_P - t) P_{vv_{C1'}}(t - t_{C1'}) P_{vv_{C2'}}(t - t_{C2'}) dt} \quad (3.12)$$

The metropolis transition kernel $p_2(x, y) = \alpha_2(x, y) q_2(x, y)$, where

$$\alpha_2(x, y) = \min \left\{ \frac{\int_{\max\{t_{C1'}, t_{C2'}\}}^{t_P} \sum_{v \in \mathcal{D}^m} P_{v_P v}(t_P - t) P_{vv_{C1'}}(t - t_{C1'}) P_{vv_{C2'}}(t - t_{C2'}) dt}{\int_{\max\{t_{C1}, t_{C2}\}}^{t_P} \sum_{v \in \mathcal{D}^m} P_{v_P v}(t_P - t) P_{vv_{C1}}(t - t_{C1}) P_{vv_{C2}}(t - t_{C2}) dt} \frac{P_{v_P v_{C3'}}(t_P - t_{C3'})(n_{T'} + 1)}{P_{v_P v_{C3}}(t_P - t_{C3})(n_T + 1)}, 1 \right\}.$$

The transition probability function Q_2 is not reversible. Thus, the probability that candidate trees are accepted into the chain, $\alpha_2(x, y)$, may be less than 1. Notice that if the C'' 's are the same as the C 's, i.e. if the rearrangement does not change the topology, then $\alpha_2(x, y) = 1$.

3.3 Properties of the algorithms

In this section we show that a Markov chain generated from $P_1(x, y)$ is uniformly ergodic, and we give a bound on the rate of convergence.

Theorem 2 *The transition probability function P_1 is uniformly ergodic.*

Proof: In (3.10) the ratio $f(t' | T'_n)/f(t^* | T_{n-1})$ is a function of the numbers of internal nodes descending from some internal nodes (see Lemma 3). Given n , those numbers are bounded, therefore the ratio is bounded. Let r and R be the lower and upper bounds of the ratio.

Consider the normalizing constant C_{x^*} in (3.10). Since this is the same as C_x that appears in (3.6), we see that C_{x^*} is the reciprocal of the sum of $2n - 3$ integrals, each of the form

$$\int_{t_k}^{t_j} \frac{\sum_{v \in \mathcal{D}^m} P_{v_j v}(t_j - t) P_{vv_k}(t - t_k) P_{vb_i}(t)}{P_{v_j v_k}(t_j - t_k)} \frac{f(\mathbf{u} | T'_n)}{f(\mathbf{t}^* | T_{n-1})} dt. \quad (3.13)$$

Since

$$\sum_{v \in \mathcal{D}^m} P_{v_j v}(t_j - t) P_{vv_k}(t - t_k) P_{vb_i}(t) \leq \sum_{v \in \mathcal{D}^m} P_{v_j v}(t_j - t) P_{vv_k}(t - t_k) = P_{v_j v_k}(t_j - t_k),$$

by the Chapman-Kolmogorov Equation, each integral of the form (3.13) is less than or equal to $\int_{t_j}^{t_k} R dt \leq R\tau$. Thus $C_{x^*} \geq ((2n - 3)R\tau)^{-1}$.

Note that when we use the graphical representation for a tree, the phylogenetic relationship is maintained whether we put the children of a node on either the right side or the left side. Thus, in the graphical representation of any tree, the branch $(-1, 1)$ can be put at the far left of the tree. Let the branch $(-1, 1)$ be at the far left in both trees x and y whose out-frames are shown in Figure 3.3. The dots on the branch $(-1, 1)$ represent a sequence of ancestral nodes of the node 1. There are at most $n - 1$ such nodes. Divide $(-1, 1)$ into $4n$ segments of equal lengths. Let $\delta = \tau/8n$. Since the total length of $(-1, 1)$ is τ , the length of each segment is 2δ . Mark segments on $(-1, 1)$ in x for the corresponding nodes in y . We will denote by S_{xy} the set of trees with the same pattern of attachments to these $4n$ segments as x and y . Note that in x , the branch $(-1, 1)$ has at least $2n$ segments that are empty, i.e. they are neither occupied nor marked.

We will establish the Doeblin condition (3.1) with $\phi = \pi$, $N_0 = 2(n - 1)$, and

$$\epsilon = \mu(S_{xy}) \left[\left(\frac{r}{(2n - 3)R\tau} \right) \left(\frac{1}{4} (1 - e^{-4\alpha\delta}) \right)^{3m} \right]^{2(n-1)}, \quad (3.14)$$

Figure 3.3: Outer frames of tree x and tree y .

where we recall that μ is defined in (3.9). We will show that with at most $2(n-1)$ transition steps, x can be transformed into y with a probability density which is greater than or equal to $\epsilon\pi(y)$, i.e. $p_1^{2(n-1)}(x, y) \geq \epsilon\pi(y)$. In the first $n-1$ transition steps, we will move the $n-1$ pendant branches in x that are connected to the external nodes $2, 3, \dots, n$ to the middle point of an empty segment on $(-1, 1)$, if they are not on $(-1, 1)$ yet. For example, if a pendant branch, say $(-i, i)$, is not on $(-1, 1)$ yet, then we can move it to the middle point of a branch, say (j, k) , which contains an empty segment. So $-i$ is at least δ away from j and k . The density of the transition is

$$\begin{aligned}
& C_x \cdot \frac{P_{v_j v_{-i}}(t_j - t_{-i}) P_{v_{-i} v_k}(t_{-i} - t_k) P_{v_{-i} b_i}(t_{-i})}{P_{v_j v_k}(t_j - t_k)} \frac{f(t | T_n)}{f(t^* | T_{n-1})} \\
& \geq \frac{r P_{v_j v_{-i}}(t_j - t_{-i})}{(2n-3)R\tau} \times P_{v_{-i} v_k}(t_{-i} - t_k) P_{v_{-i} b_i}(t_{-i}).
\end{aligned} \tag{3.15}$$

The contributions of each site to $P_{v_j v_{-i}}(t_j - t_{-i})$, $P_{v_{-i} v_k}(t_{-i} - t_k)$, and $P_{v_{-i} b_i}(t_{-i})$ are all greater than or equal to $(1 - e^{-4\alpha\delta})/4$ (see 2.7). Using this fact in conjunction with inequality (3.15), we see that

$$C_x \frac{P_{v_i v}(t_i - t) P_{v v_j}(t - t_j) P_{v b_k}(t)}{P_{v_i v_j}(t_i - t_j)} \frac{f(\mathbf{t} \mid T_n)}{f(\mathbf{t}^* \mid T_{n-1})} \geq \frac{r}{(2n-3)R\tau} \left[\frac{1}{4}(1 - e^{-4\alpha\delta}) \right]^{3m} \stackrel{\text{set}}{=} M_1.$$

Let x' be the tree after the first $n-1$ transitions. With probability greater than or equal to M_1^{n-1} every pendant branch of x' is attached to $(-1, 1)$. Thus we have shown that $p_1^{n-1}(x, x') \geq M_1^{n-1}$ and $p_1^{n-1}(y, x') \geq M_1^{n-1}$.

Recall that the chain Q_1 is reversible, so that $p_1(x, y) = q_1(x, y)$. Thus P_1 is reversible, and we have also $\pi(u)p_1^k(u, v) = \pi(v)p_1^k(v, u)$ for any k . Hence,

$$\begin{aligned} p_1^{2(n-1)}(x, y) &\geq \int_{x' \in S_{xy}} p_1^{n-1}(x, x') p_1^{n-1}(x', y) d\mu(x') \\ &\geq M_1^{n-1} \int_{x' \in S_{xy}} p_1^{n-1}(x', y) d\mu(x') \\ &= M_1^{n-1} \int_{x' \in S_{xy}} p_1^{n-1}(y, x') \frac{\pi(y)}{\pi(x')} d\mu(x') \\ &\geq M_1^{2(n-1)} \pi(y) \int_{x' \in S_{xy}} \frac{1}{\pi(x')} d\mu(x') \\ &\geq \mu(S_{xy}) M_1^{2(n-1)} \pi(y) = \epsilon \pi(y), \end{aligned}$$

where the last inequality follows from Jensen's inequality. This completes the proof of Theorem 2. ■

The convergence of the chain with transition probability function P_2 follows from the remark after Lemma 2 and Theorem 1 of Athreya, Doss, and Sethuraman (1996).

3.4 Computer implementation

Although our model and two Metropolis-Hastings samplers are applicable to any complex molecular model (for example the five-parameter model), to illustrate the methodology, we begin our study with the one-parameter model. Its possible extension will be discussed in Chapter 6.

In order to implement our algorithms efficiently, we use a computer friendly representation of phylogenetic trees. Notice that a phylogenetic tree represents an ancestry-descent relationship. Each node in a phylogenetic tree has either two children or no children at all. So if we list parent nodes on the first row and their two children nodes on the second and third rows under the same column, this three row labeling represents a topology. This representation is called the three-row representation. Furthermore if each node in a three-row representation is associated with a nucleotide sequence and a splitting time, it will represent a tree. We still label external nodes with positive numbers $1, 2, \dots, n$. But we label internal nodes with positive numbers after n , $n+1, n+2, \dots, 2n-1$, for convenience in indexes. For example, the three-row representation of the topology in Figure 1.1(a) is

$$\begin{pmatrix} 6 & 7 & 8 & 9 \\ 1 & 8 & 4 & 2 \\ 7 & 9 & 5 & 3 \end{pmatrix}.$$

Advantages of a three-row representation are (i) it is computationally efficient in terms of memory; and (ii) three-row representations of two topologies, before and after a transition (either in the pendant branch strategy or in the local rearrangement strategy), are different only on two columns.

In a Markov chain generating process, we keep tracking unique topologies and cumulating their frequencies. When each new tree is generated, we compare its topology

with all previously saved topologies. If the topology repeats one of the saved topologies, we increment the cumulative frequency of that topology by one. Otherwise we need to save the new topology. However the three-row representation is not unique; each topology can be represented by more than one three-row representation. This property of the three-row representation makes a comparison between topologies hard. We must have another representation which is not only computer efficient but which also uniquely describes a topology to utilize the comparisons. One candidate for such a representation is the generational matrix whose $(i, j)^{th}$ entry is the generation at which taxon i and taxon j split. For example, the symmetric generational matrix which uniquely describes the tree in Figure 1.1(a) is

$$\begin{pmatrix} - & 1 & 1 & 1 & 1 \\ 1 & - & 3 & 2 & 2 \\ 1 & 3 & - & 2 & 2 \\ 1 & 2 & 2 & - & 3 \\ 1 & 2 & 2 & 3 & - \end{pmatrix}.$$

Although a generational representation uniquely represents a topology, an ancestry-descent relationship in a generational representation is not as clear as in a three-row representation. Thus the two representations must be used in a complementary manner. The three-row representation is used to do transitions from the current tree to the next tree. After the next tree is generated, the three-row representation of the tree is converted to its generational representation and compared with all previously saved generational representations of the unique trees. For each unique tree only lower diagonal entries of its generational representation are saved, because they are sufficient to represent a topology. The generational representation is also used to compute the triples distance between two trees (Critchlow, Pearl, and Qian 1994), which will be discussed in Chapter 4.

In our probability model we assume that the mutation rate α and the splitting time at the root τ are two unknown parameters. In the implementation of the algorithms we estimate these parameters during the Markov chain generation process. The estimations of α and τ are also important in phylogenetic studies. In our probability model the probability of tree depends upon the parameters α and τ only through their product $\alpha\tau$. So we can only estimate the product $\alpha\tau$ in our model. Without loss of generality we set $\tau = 1$ and we estimate the parameter α .

Note that from (2.7), the unconditional probability that two sequences are different at a particular site at time t is $p = (3/4)(1 - \exp(-8t\alpha))$, so that

$$\alpha t = -(1/8) \ln(1 - 4p/3). \quad (3.16)$$

Replacing p by $\hat{p}$, the observed proportion of different nucleotides between two sequences, we get an estimate of the mutation rate per site between the two sequences. Since all taxa evolved from a common ancestor and we have assumed a constant mutation rate over the period of evolution, we may take the average of estimates based on all possible pairs of the sequences as our initial estimate of α , i.e.

$$\frac{2}{n(n-1)} \sum_{1 \leq i < j \leq n} -\frac{1}{8} \ln(1 - 4\hat{p}_{ij}/3),$$

where $\hat{p}_{ij}$ is the proportion of different nucleotides between the sequence of the i^{th} taxon and the sequence of the j^{th} taxon. If there is a prespecified outgroup taxon, then the initial estimate of α is

$$\frac{1}{n} \sum_i^n -\frac{1}{8} \ln(1 - 4\hat{p}_i/3),$$

where $\hat{p}_i$ is the proportion of different nucleotides between the sequences of the i^{th} taxon and the sequence of the outgroup. (The initial estimate of α is used during the

generation of the initial tree.) Our estimate of α is then updated during the burn-in process of the Markov chain generation.

Once a tree is generated, we know the branching pattern, branch lengths, and nucleotide sequences of all nodes (external nodes as well as internal nodes). Let i index all branches, so that i ranges from 1 to $2n$, and let l_i be the length of the i^{th} branch. Using (2.7) again, we have

$$\alpha l_i = -(1/4) \ln(1 - 4p_{[l_i]}/3),$$

where $p_{[l_i]}$ is the probability that for branch l_i , a single site in the parent node will have a different nucleotide than the child node. Summing and dividing by $\sum_{i=1}^{2n} l_i$, we get an expression for α , which suggests the estimate

$$\hat{\alpha} = \frac{\sum_{i=1}^{2n} -\frac{1}{4} \ln(1 - 4\hat{p}_{[l_i]}/3)}{\sum_{i=1}^{2n} l_i}, \quad (3.17)$$

where $\hat{p}_{[l_i]}$ is the observed proportion of different nucleotides between the sequences of the two nodes which are connected by the i^{th} branch. (Certainly, other estimates of α are also possible.) The average of this estimate and the estimate from all previous burn-in trees is taken as the mutation rate for the next tree generation. We repeat the updating in the first half of the burn-ins. In the second half of the burn-ins we use the estimate from the first half as our initial estimate of α and repeat the update process in order to reduce the influence of the first initial estimate. The estimate at the end of the second half of the burn-ins is then fixed in the rest of the tree generation processing.

A Markov chain generation starts at an initial tree which is generated using a clustering algorithm based on pairwise distances. Although qualitatively, the convergence

of the Markov chain does not depend on the initial state, a good initial tree makes the chain converge faster. Since we estimate the mutation rate per site during the burn-in process, a good initial tree helps us to get a more accurate estimate within a shorter period of burn-ins. This is especially true since with either of our transitional processes, each transition allows only one branch to be altered. Changes are gradual. If we start with a tree which is less likely to explain the observed sequences of taxa, the burn-in time required will be longer.

Our modified UPGMA algorithm produces a topology which may not be the most probable tree, but in practice it provides a good starting point for the Markov chain. Like any distance related method we first compute the pairwise distances of all taxa. Then draw a pair of nodes to be clustered first with probability which is proportional to $(\frac{p}{1-p})^d$ where p is the probability that one nucleotide mutates to the other three nucleotides during τ period of time and d is the number of different nucleotides between sequences of that pair. After a pair of sequences are chosen to be clustered, a nucleotide sequence is generated by randomly drawing between the two sequences at each site. Then this newly generated sequence and the remaining sequences which have not yet been clustered are used to create a new distance matrix. This time the number of sequences has been reduced by one. We repeat the clustering process until only two sequences are left. This whole process generates a topology. Next we generate the splitting times and nucleotide sequences of internal nodes.

Given the topology of a tree, the splitting times of internal nodes are generated according to the conditional distribution of times given a topology that was defined in Chapter 2. As was assumed in Chapter 2, the time of the root is a parameter, τ . Given the time of the root, the times of its two children are independently distributed

with a cumulative distribution function, $F(t|\tau) = (\frac{t}{\tau})^{n_{\text{descent}}+1}$, where n_{descent} is the number of internal nodes which are descended from the node. Thus given τ , the time of a child of the root, t , is picked by taking $t = u^{\frac{1}{n_{\text{descent}}+1}}\tau$ where $u \sim U(0, 1)$. Similarly we pick the times of the internal nodes on the third generation, fourth generation and etc.

Having a known topology and time of tree, finally we generate the sequences of the internal nodes. However unlike the generations of times, the generations of nucleotide sequences for the internal nodes start from the parents of the external nodes, because the sequences of external nodes are known. First we generate nucleotide sequences for the internal nodes whose two children are external nodes. For example, the internal node k has two children, i and j , which are external nodes. The nucleotide sequence of k at a site is generated according to the probability

$$\frac{P_{v_k v_i}(t_k - t_i)P_{v_k v_j}(t_k - t_j)}{\sum_{v \in \mathcal{D}} P_{v v_i}(t_k - t_i)P_{v v_j}(t_k - t_j)} = \frac{P_{v_k v_i}(t_k - t_i)P_{v_k v_j}(t_k - t_j)}{P_{v_i v_j}(2t_k - t_i - t_j)}.$$

After the nucleotide sequences of the internal nodes at that generation have been generated, we repeat this process for the next generation of internal nodes so that the sequences of their two children have already been generated. We process in this manner until the sequence of the root node is generated.

During a Markov chain generation, we often need to compute a summation like

$$\sum_{v \in \mathcal{D}^m} P_{v_P v}(t_P - t)P_{v v_{C_1}}(t - t_{C_1})P_{v v_{C_2}}(t - t_{C_2}) \quad (3.18)$$

where (v_P, t_P) , (v_{C_1}, t_{C_1}) and (v_{C_2}, t_{C_2}) are nucleotide sequences and times of parent node, child #1, and child #2 of a working node. A working node is the attaching point in the algorithm P_1 or the target node in the algorithm P_2 . This summation depends upon the sequences of the parent node and the two children nodes. Since nucleotide

substitutions are independent across sites, the summation in (3.18) is a product of m such summations at each site. At each site, the dependence of the summation on the sequences of the parent and the two children could be different based on the different molecular evolutionary model. For example, for the one-parameter model, the summation has five possible forms based on five distinct relationships among nucleotides of parent and two children at a site, namely

- (i) $v_P = v_{C_1} = v_{C_2}$,
- (ii) v_P , v_{C_1} , and v_{C_2} are pairwise different,
- (iii) v_{C_1} and v_{C_2} are the same but different from v_P ,
- (iv) v_P and v_{C_1} are the same but different from v_{C_2} , and
- (iv) v_P and v_{C_2} are the same but different from v_{C_1} .

The five corresponding summations after some numerical manipulation become

- (i) $\frac{1}{4^2} \left(1 + 3e^{-4\alpha(t_P - t_{C_1})} + 12e^{-4\alpha(t_P - t_{C_2})} + 12e^{-4\alpha(2t - t_{C_1} - t_{C_2})} + 24e^{-4\alpha(t_P + t - t_{C_1} - t_{C_2})} \right)$,
- (ii) $\frac{1}{4^2} \left(1 - e^{-4\alpha(t_P - t_{C_1})} - e^{-4\alpha(t_P - t_{C_2})} - e^{-4\alpha(2t - t_{C_1} - t_{C_2})} + 2e^{-4\alpha(t_P + t - t_{C_1} - t_{C_2})} \right)$,
- (iii) $\frac{1}{4^2} \left(1 - e^{-4\alpha(t_P - t_{C_1})} - e^{-4\alpha(t_P - t_{C_2})} + 3e^{-4\alpha(2t - t_{C_1} - t_{C_2})} - 2e^{-4\alpha(t_P + t - t_{C_1} - t_{C_2})} \right)$,
- (iv) $\frac{1}{4^2} \left(1 + 3e^{-4\alpha(t_P - t_{C_1})} - e^{-4\alpha(t_P - t_{C_2})} - e^{-4\alpha(2t - t_{C_1} - t_{C_2})} - 2e^{-4\alpha(t_P + t - t_{C_1} - t_{C_2})} \right)$, and
- (v) $\frac{1}{4^2} \left(1 - e^{-4\alpha(t_P - t_{C_1})} + 3e^{-4\alpha(t_P - t_{C_2})} - e^{-4\alpha(2t - t_{C_1} - t_{C_2})} - 2e^{-4\alpha(t_P + t - t_{C_1} - t_{C_2})} \right)$,

where t is the time of the working node.

Integration of the summation in Equation (3.18) is computed based on Romberg's method. The C program of Romberg's integration is given in Press et al. (1992).

The amount of computer CPU time needed to carry out the algorithm based on P_1 grows linearly with n for a fixed chain length. This makes P_1 inefficient in problems with a large number of taxa. This same drawback is found in the algorithm proposed by Mau, Newton, and Larget (1996) who report dramatically increased CPU time and other computational difficulties arising from the leap to a large number of taxa.

The Markov chain P_2 avoids the “curse of dimensionality” that hinders the other proposals. This is because the calculation of the transition density involves only the nodes and branches around the local rearrangement (see (3.12)). Thus, for a fixed chain length, the CPU time needed does not grow even linearly with n (although for any of the proposed algorithms, the length of the chain necessary to ensure convergence will grow with n). For example, the CPU time in running the chain of 1 million trees and 10,000 burn-ins on nine Bacteriophage data in Chapter 5 was 67 minutes; and the CPU time in running the chain of 1 million trees and 10,000 burn-ins on 30 simulated sequences with 500 nucleotide sites was about one hour. The chains were generated using a HP workstation doing about 24 million floating point operations per second.

CHAPTER 4

HOW WELL DOES THE MARKOV CHAIN MONTE CARLO METHOD WORK?

In Chapter 3 we gave two Metropolis-Hastings transition probabilities and discussed their computer implementation for the one-parameter model. We also proved that the Markov chains generated from both algorithms converge to the conditional probability distribution of trees given observable external sequences. How well do the algorithms actually work? To answer this question, in this chapter we first compare the probabilities of topologies estimated from our two Markov chain Monte Carlo algorithms with true probabilities of topologies calculated directly from the conditional probability of topologies defined in (2.11) in the case when $n = 3$. When direct calculation of the conditional probability from the formula becomes hard and impossible, we use two diagnostic tests to assess convergence of the chain for any size tree. Also in this chapter, we assess the reliability of subtrees and propose three estimates of the true phylogeny and measures of the variability of the estimates.

4.1 Comparison with true probabilities of topologies that consist of three in-group taxa and one outgroup

The minimum tree space for algorithm P_1 is a space of trees with three in-group taxa and one outgroup taxon. There are three distinct topologies in this space shown

in Figure 4.1. Now we describe how the true conditional probabilities of these three topologies are computed.

Figure 4.1: Three possible topologies in $\mathcal{T}_3$.

If the one-parameter molecular evolutionary model is assumed, the nucleotide sequences of the four taxa at a single site affect the probabilities of topologies through the following 15 possible ways:

1. $b_1 = b_2 = b_3 = b_0$,
2. $b_1 = b_2 = b_3 \neq b_0$,
3. $b_i, i = 0, 1, 2, 3$ are different,
4. b_1, b_2, b_3 are different and $b_0 = b_1$,
5. b_1, b_2, b_3 are different and $b_0 = b_2$,
6. b_1, b_2, b_3 are different and $b_0 = b_3$,
7. b_0, b_1, b_3 are different and $b_1 = b_2$,

8. $b_1 = b_2 = b_0 \neq b_3$,
9. $b_1 = b_2 \neq b_3 = b_0$,
10. b_0, b_1, b_2 are different and $b_1 = b_3$,
11. $b_1 = b_3 = b_0 \neq b_2$,
12. $b_1 = b_3 \neq b_2 = b_0$,
13. b_0, b_1, b_2 are different and $b_2 = b_3$,
14. $b_2 = b_3 = b_0 \neq b_1$, and
15. $b_2 = b_3 \neq b_1 = b_0$.

The term $\sum_{\mathbf{v} \in \mathcal{P}^{3m}} P(T_1, \mathbf{t}, \mathbf{v}, \mathbf{b})$, which is a part of the probability of being T_1 defined in (2.11), can be derived for each corresponding category which describes a relationship among nucleotides of the four external sequences at a site. The results are denoted by $S_i(\mathbf{t})$ for $i = 1, \dots, 15$.

$$S_1(\mathbf{t}) = \frac{1}{4^3} \left(1 + 9e^{-8\alpha\tau} + 3e^{-8\alpha t_{-3}} + 9e^{-8\alpha(\tau+t_{-3})} + 6e^{-8\alpha t_{-2}} + 12e^{-4\alpha(2\tau+t_{-2})} \right. \\ \left. + 6e^{-4\alpha(2t_{-2}+t_{-3})} + 6e^{-4\alpha(2\tau+t_{-3})} + 12e^{-4\alpha(2\tau+t_{-2}+t_{-3})} \right),$$

$$S_2(\mathbf{t}) = \frac{1}{4^3} \left(1 - 3e^{-8\alpha\tau} + 3e^{-8\alpha t_{-3}} - 3e^{-8\alpha(\tau+t_{-3})} + 6e^{-8\alpha t_{-2}} - 4e^{-4\alpha(2\tau+t_{-2})} \right. \\ \left. + 6e^{-4\alpha(2t_{-2}+t_{-3})} - 2e^{-4\alpha(2\tau+t_{-3})} - 4e^{-4\alpha(2\tau+t_{-2}+t_{-3})} \right),$$

$$S_3(\mathbf{t}) = \frac{1}{4^3} \left(1 - 3e^{-8\alpha\tau} - e^{-8\alpha t_{-3}} + e^{-8\alpha(\tau+t_{-3})} - 2e^{-8\alpha t_{-2}} + 4e^{-4\alpha(2\tau+t_{-2})} \right. \\ \left. + 2e^{-4\alpha(2t_{-2}+t_{-3})} + 2e^{-4\alpha(2\tau+t_{-3})} - 4e^{-4\alpha(2\tau+t_{-2}+t_{-3})} \right),$$

$$S_4(t) = S_5(t) = \frac{1}{4^3} \left(1 + e^{-8\alpha\tau} - e^{-8\alpha t_{-3}} + e^{-8\alpha(\tau+t_{-3})} - 2e^{-8\alpha t_{-2}} + 2e^{-4\alpha(2t_{-2}+t_{-3})} - 2e^{-4\alpha(2\tau+t_{-3})} \right),$$

$$S_6(t) = \frac{1}{4^3} \left(1 + e^{-8\alpha\tau} - e^{-8\alpha t_{-3}} - 3e^{-8\alpha(\tau+t_{-3})} - 2e^{-8\alpha t_{-2}} - 4e^{-4\alpha(2\tau+t_{-2})} + 2e^{-4\alpha(2t_{-2}+t_{-3})} + 2e^{-4\alpha(2\tau+t_{-3})} + 4e^{-4\alpha(2\tau+t_{-2}+t_{-3})} \right),$$

$$S_7(t) = \frac{1}{4^3} \left(1 - 3e^{-8\alpha\tau} + 3e^{-8\alpha t_{-3}} - 3e^{-8\alpha(\tau+t_{-3})} - 2e^{-8\alpha t_{-2}} + 4e^{-4\alpha(2\tau+t_{-2})} - 2e^{-4\alpha(2t_{-2}+t_{-3})} - 2e^{-4\alpha(2\tau+t_{-3})} + 4e^{-4\alpha(2\tau+t_{-2}+t_{-3})} \right),$$

$$S_8(t) = \frac{1}{4^3} \left(1 + 5e^{-8\alpha\tau} + 3e^{-8\alpha t_{-3}} - 3e^{-8\alpha(\tau+t_{-3})} - 2e^{-8\alpha t_{-2}} + 4e^{-4\alpha(2\tau+t_{-2})} - 2e^{-4\alpha(2t_{-2}+t_{-3})} + 6e^{-4\alpha(2\tau+t_{-3})} - 4e^{-4\alpha(2\tau+t_{-2}+t_{-3})} \right),$$

$$S_9(t) = \frac{1}{4^3} \left(1 + e^{-8\alpha\tau} + 3e^{-8\alpha t_{-3}} + 9e^{-8\alpha(\tau+t_{-3})} - 2e^{-8\alpha t_{-2}} - 4e^{-4\alpha(2\tau+t_{-2})} - 2e^{-4\alpha(2t_{-2}+t_{-3})} - 2e^{-4\alpha(2\tau+t_{-3})} - 4e^{-4\alpha(2\tau+t_{-2}+t_{-3})} \right),$$

$$S_{10}(t) = S_{13}(t) = \frac{1}{4^3} \left(1 - 3e^{-8\alpha\tau} - e^{-8\alpha t_{-3}} + e^{-8\alpha(\tau+t_{-3})} + 2e^{-8\alpha t_{-2}} - 2e^{-4\alpha(2t_{-2}+t_{-3})} + 2e^{-4\alpha(2\tau+t_{-3})} \right),$$

$$S_{11}(t) = S_{14}(t) = \frac{1}{4^3} \left(1 + 5e^{-8\alpha\tau} - e^{-8\alpha t_{-3}} - 3e^{-8\alpha(\tau+t_{-3})} + 2e^{-8\alpha t_{-2}} + 4e^{-4\alpha(2\tau+t_{-2})} - 2e^{-4\alpha(2t_{-2}+t_{-3})} - 2e^{-4\alpha(2\tau+t_{-3})} - 4e^{-4\alpha(2\tau+t_{-2}+t_{-3})} \right), \text{ and}$$

$$S_{12}(t) = S_{15}(t) = \frac{1}{4^3} \left(1 + e^{-8\alpha\tau} - e^{-8\alpha t_{-3}} + e^{-8\alpha(\tau+t_{-3})} + 2e^{-8\alpha t_{-2}} - 4e^{-4\alpha(2\tau+t_{-2})} - 2e^{-4\alpha(2t_{-2}+t_{-3})} - 2e^{-4\alpha(2\tau+t_{-3})} + 4e^{-4\alpha(2\tau+t_{-2}+t_{-3})} \right).$$

Suppose that there are m_i sites for which the nucleotides of the four sequences satisfy the relationship listed in the i^{th} category for $i = 1, \dots, 15$. The probability of being

the first topology in Figure 4.1 given the nucleotide sequences of the external nodes is then

$$P(T_1|\mathbf{b}) = C \int_0^\tau \int_0^{t-2} \prod_{i=1}^{i=3} (S_i(t))^{m_i} (S_4(t))^{m_4+m_5} \prod_{i=6}^{i=9} (S_i(t))^{m_i} (S_{10}(t))^{m_{10}+m_{13}} \\ (S_{11}(t))^{m_{11}+m_{14}} (S_{12}(t))^{m_{12}+m_{15}} f(t|T_1) dt_{-3} dt_{-2} \quad (4.1)$$

where C is a normalizing constant.

The second topology in Figure 4.1 becomes the first topology if we swap the positions of the second and the third sequences. The number of sites which correspond to the fifteen S 's are $m_i, i = 1, 2, 3, m_4, m_6, m_5, m_j, j = 10, 11, 12, m_k, k = 7, 8, 9, m_l, l = 13, 14, 15$. So the probability of being the second topology in Figure 4.1 given nucleotide sequences of external nodes is

$$P(T_2|\mathbf{b}) = C \int_0^\tau \int_0^{t-2} \prod_{i=1}^{i=3} (S_i(t))^{m_i} (S_4(t))^{m_4+m_6} (S_6(t))^{m_5} (S_7(t))^{m_{10}} (S_8(t))^{m_{11}} \\ (S_9(t))^{m_{12}} (S_{10}(t))^{m_7+m_{13}} (S_{11}(t))^{m_8+m_{14}} (S_{12}(t))^{m_9+m_{15}} f(t|T_2) dt_{-3} dt_{-2}. \quad (4.2)$$

Similarly the third topology in Figure 4.1 becomes the first topology if we swap the positions of the first and third sequences. The number of sites which correspond to the fifteen S 's are $m_i, i = 1, 2, 3, m_5, m_6, m_4, m_j, j = 13, 14, 15, m_k, k = 7, 8, 9, m_l, l = 10, 11, 12$. So the probability of being the third topology in Figure 4.1 given nucleotide sequences of external nodes is

$$P(T_3|\mathbf{b}) = C \int_0^\tau \int_0^{t-2} \prod_{i=1}^{i=3} (S_i(t))^{m_i} (S_4(t))^{m_5+m_6} (S_6(t))^{m_4} (S_7(t))^{m_{13}} (S_8(t))^{m_{14}} \\ (S_9(t))^{m_{15}} (S_{10}(t))^{m_7+m_{10}} (S_{11}(t))^{m_8+m_{11}} (S_{12}(t))^{m_9+m_{12}} f(t|T_3) dt_{-3} dt_{-2}. \quad (4.3)$$

0	ACGCATCGCATCTGGCACGAGCCAGTGACGCCTAGAATGCAGGTTTGTCTTTCCAGGGGATATTGAAGAC
1	-----A-----A---T-----T-----C-----
2	-----A-----
3	-----G-----
0	TATGACACGCGGGCAACGCATCAAGTGGACCCCTGATAACCTAAGAAACCAGGACCATAGCTTGCGCGCA
1	-----
2	-----
3	-----
0	ACAAAGCCTTGGGCTATCGAGATCCGAATTGTGGTGCAGGATCGGCGGGCGCTGGGCCCC
1	-----A-----
2	-----A-----
3	-----C-A-----

Table 4.1: Simulated four sequences with 200 sites for each. The numbers on the left column are the labels of the sequences. The sequence labeled by 0 is the outgroup to the sequences 1-3. The nucleotides in the sequences 1-3 which are identical to the nucleotides in the sequence 0 are indicated by dashes; those which are different from the sequence 0 are represented by letters.

Finally we can find C since $P(T_1|\mathbf{b}) + P(T_2|\mathbf{b}) + P(T_3|\mathbf{b}) = 1$.

To compare the estimated probabilities of these three topologies using the two algorithms P_1 and P_2 with the probabilities calculated using (4.1), (4.2), and (4.3), we simulated four sequences, one outgroup taxon and three in-group taxa, using the one-parameter model with $\alpha = 0.005$ (shown in Table 4.1). We assumed that a tree, which the outgroup 0 was split from the three in-group taxa at the root while the third taxon was split from the first and second taxa after one-third period of evolutionary time and the second taxon was split from the first taxon after another one-third period of evolutionary time, was the true tree. The sequences of four taxa were generated by simulating evolutionary changes along the tree. We also assumed that there were 200 nucleotide sites. At each site nucleotides evolved independently, starting from a random nucleotide sampled equiprobably from A, G, C, and T and simulating change according to the probability in Equation 2.7.

The exact probabilities of the three topologies from expressions (4.1), (4.2), and (4.3) were 0.924, 0.021, and 0.055. Using P_1 we generated 20,000 trees and discarded the first 10,000. Estimation of the probabilities of three topologies from the last 10,000 trees gave 0.928, 0.019, and 0.054, respectively. The estimate of the mutation rate per site, $\widehat{\alpha\tau}$, during the burn-in period provided by the first 10,000 trees of this generating process was 0.004. Similarly, chain P_2 produced the estimated probabilities 0.920, 0.022, and 0.058, with $\widehat{\alpha\tau} = 0.004$. Our experience also indicated that the algorithms were reasonably robust to small changes in the estimate of α . For instance, when we ran the two chains using the true $\alpha = .005$, P_1 estimated the three tree probabilities to be 0.928, 0.020, and 0.052, while P_2 gave estimates of 0.925, 0.021, and 0.054.

4.2 Diagnostics

One important issue in Markov chain Monte Carlo simulations is to assure that the generated chain converges to the target distribution. Many diagnostic tools have been proposed in the Markov chain Monte Carlo literature (for a review see for example Kathryn, Cowles and Carlin 1995). Since our tree space is so different from the ordinary real valued space, most of these diagnostics are not applicable in our case. However, we will use one existing diagnostic, namely the scaled regeneration quantile plot (see Mykland, Tierney and Yu 1995), as a primary test for convergence of a chain. Then we will give a new diagnostic. For this diagnostic, we generate two chains starting from two different initial trees. The diagnostic is designed to test whether the two chains converge to the same distribution. If the two chains do not converge to the same distribution, it indicates that there is at least one chain which

does not converge. If that is the situation, it is probably necessary to have a longer burn-in process.

4.2.1 Scaled regeneration quantile (SRQ) plot

A stochastic process $X_n : n = 0, 1, \dots$ is regenerative if there are times $T_0 \leq T_1 \leq \dots$ such that at each T_i , the future of the process is independent of the past and identically distributed. Then the tours of the process between these times are iid, and the times themselves form a renewal process. The renewal process is delayed if $T_0 \neq 0$.

A *Scaled Regenerative Quantile* (SRQ) plot is a plot of T_i/T_n against i/n . An SRQ plot can provide a useful diagnostic for examining the performance of a Markov chain sampler. If the chain length is long enough, this plot should be close to a straight line through the origin with unit slope. Deviations from this line suggest that the burn-in process is not long enough for the sampler to reach equilibrium.

The key issue of applying this diagnostic is to identify regeneration times. Generally regeneration times can be obtained by observing times that the chain returns to a pre-specified set, A , which satisfies $\pi(A) > 0$. If the chain converges to π , $T_i - T_{i-1}$ should be independently identically distributed. In the present case, a topology or a few topologies are chosen as the set A . A topology is a set in which any tree shares the same topology. Choosing an appropriate size for the set A is the most difficult task in our case. We need to pick this set A prior to the generation of chain. Generally we do not have any knowledge about the distribution of topologies when we are given a group of sequences. The size of A affects the quality of the SRQ plot. Therefore it affects the conclusion using the SRQ plot as a diagnostic. If A is too big, the chain

hardly leaves the set A . Too many regeneration times are generated to detect divergence within the set A . Also the $T_i - T_{i-1}$ may no longer be iid. Conversely if A is too small, the chain hardly goes into A . Therefore we will not get enough regeneration times to do a SRQ plot. This is the main drawback of the SRQ plot diagnostic.

4.2.2 Inter-value distance comparisons

In this subsection we introduce a new diagnostic test. For this diagnostic we do not need to have any prior knowledge about the distribution of topologies. But we need to generate two chains. This diagnostic is based on the fact that two chains should converge to the same distribution, namely the conditional distribution of topologies given the observed sequences of the external nodes.

To see if two distributions on the space of trees are identical we appeal to the theorem in Maa, Pearl, and Bartoszynski (1996). This theorem states that for independent replicates X_1, X_2, X_3 from distribution F and independent replicates Y_1, Y_2, Y_3 from distribution G , then $F = G$ if and only if the one-dimensional distances $\rho(X_1, X_2)$, $\rho(Y_1, Y_2)$, and $\rho(X_3, Y_3)$ all have the same distributions. (Here, $\rho(a, b)$ is an arbitrary distance between a and b .) In the present case, since we are focusing on the distribution of tree topologies, we choose the triples distance (Critchlow, Pearl, and Qian 1996) between trees as a distance function. The triples distance between two rooted trees is the proportion of the number of rooted subtrees of size three that are different in the two trees out of all possible rooted subtrees of size three. The triples distance provides a measure of the discrepancy between two trees by reporting the proportion of triples of taxa whose relationship differs in the two trees. Consider the issue of whether or not two chains generated from the transition probability function

P have converged. To apply the theorem, we randomly draw 10,000 pairs of topologies (one from each chain) and calculate the triples distances for those pairs. And we also randomly draw 10,000 pairs of topologies from within each chain and calculate their triples distances. Then we draw the histograms of these 10,000 distances. Discrepancies among these three histograms suggest a difference between the two chains. Therefore they indicate that one or both chains did not converge to the appropriate conditional distribution.

A formal significance test of the equality of the two distributions can be conducted using the observed distances (e.g. by adapting the 3-sample test of Kiefer 1959 or by picking a smaller number of trees systematically and applying the test based on minimal spanning trees of Friedman and Rafsky 1981).

4.3 Reliability of subtrees

Given observed DNA sequences of taxa, molecular geneticists seek to find an estimate of phylogeny and quantify the degree of uncertainty in that estimate. Other competitive tree methods (reviewed in Chapter 1) give only one or a few phylogenies without indicating the level of variability of the estimate. The bootstrap method is usually used to ascertain the uncertainty in the estimate. In the bootstrap, the data points are sampled randomly with replacement from the original data set until a new data set containing the original number of observations is obtained. The bootstrap method is used to infer a confidence interval of a group taxa as monophyly. A group is monophyletic on a tree if and only if it contains all of the descendants of its most recent common ancestor. For confidence limits to be valid, the bootstrap assumes that the data at hand are representative of the underlying distribution and that the

reconstruction method makes consistent estimate of the phylogeny. Any violation of these two assumptions makes the bootstrap confidence limits difficult to interpret.

A second approach (see for example Cracraft and Helm-Bychowski 1991) examines sets of near-most-parsimonious trees to see at what degree a particular cladistic structure is corroborated across those trees. A drawback of this approach is that near-most-parsimonious trees do not cast the whole distribution of trees inferred from the given data set.

The Markov chain Monte Carlo method offers an alternative approach. Markov chains generated using the Markov chain Monte Carlo method has been proved to converge to the conditional probability of trees provided model assumptions are correct. Thus using the generated trees, we can correctly estimate a confidence limit for any aspect of interest for the phylogenetic relationship which includes getting confidence levels regarding whether a group of taxa form a monophyly.

Generally a consensus tree is used to summarize areas of agreement among bootstrap trees or near-most-parsimonious trees. Of the existing consensus tree methods, the strict consensus (Schuh and Polhemus 1981, Sokal and Rohlf 1981) was defined as the unique tree that contains only those groups that appear on all of the rival trees. This method is conceptually the simplest but is too strict. Because two trees that support identical relationships for all but one of the taxa may share no groupings when the last taxon is included, it is possible that a strict consensus tree is a completely unresolved bush. A majority rule consensus method (Margush and McMorris 1981) relaxes the rule of agreement among 100% of the trees to any prespecified proportion of the trees. Typically, 50% is used, so that all groups occurring on more than half of the trees are retained in the "50% majority rule" consensus tree. Sometimes the

prespecified proportion goes below 50% in order to resolve a tree completely. The consensus tree is represented by a tree with percentages labeling each fork indicating the percentage of trees that duplicate the external nodes forming the descendant subtree. The chain generated using the Markov chain Monte Carlo method can directly be used to create a consensus tree.

4.4 Estimate of the true phylogeny

In this subsection we propose three estimates of the true phylogeny and the measures of their variability based on the chain generated using the Markov chain Monte Carlo method.

In our first proposal, we choose the modal tree as an estimate of the true phylogeny, i.e. the tree that occurs most frequently in the chain is chosen as an estimate of the true phylogeny. In our second proposal, we choose the consensus tree as an estimate of the true phylogeny. In our last proposal, we choose the tree which is closest to a majority of trees in the chain as an estimate of the true phylogeny. To find the estimate in the third proposal, for each topology in the chain, we calculate triples distances between this topology and the other topologies in the chain and find the minimum distance needed to cover 90% (or 95%) of the topologies. We call this minimum distance the radius. The topology with the least radius is the topology closest to 90% (or 95%) of the topologies, therefore it is the estimate of the true phylogeny in our third proposal. The variability of each estimate is measured by the radius which covers 90% (or 95%) of the topologies using the estimate as the center. The larger the radius is, the higher the degree of uncertainty in the estimate. It also means that the sequences of taxa carry less information about the true phylogeny.

Conversely a small radius indicates less uncertainty in the estimate and that the sequences of taxa are more informative in constructing the true phylogeny. The trees that are covered by the least radius compose a 90% (or 95%) credible region for the true phylogeny.

In the example used in Section 4.1 the estimated probability of T_1 in Figure 4.1 is 0.928 using P_1 and 0.920 using P_2 . All three proposals give the same estimate, T_1 . The radius to cover 90% of the topologies is 0. So 90% credible region includes only one topology, i.e. T_1 . This is an example in which the given sequences are very informative for inferring the true topology. In another example, suppose that the given sequences have no informative sites. The probabilities of the three topologies will be equal. Estimates of the true topology could be any one of three topologies. The minimum radius is one for any confidence level over 100/3%.

CHAPTER 5

APPLICATIONS

In this chapter we apply the Markov chain Monte Carlo method to three data sets: nine Bacteriophage T7 experimental data, 16 species of fish data, and 30 viral sequences data from Russian children with HIV, to demonstrate the usage of our algorithms in real data sets.

5.1 The Bacteriophage T7 data

To illustrate our algorithm, first we return to the example mentioned in Chapter 1, in which Hillis et al. (1992) used a system based on serial propagation of Bacteriophage T7 in the presence of a mutagen to create the first completely known phylogeny. Figure 5.1(a) shows the true phylogeny.

The eight in-group taxa studied by Hillis et al. consisted of sequences with 1091 sites each. How much information about the true phylogeny is contained in these eight sequences? To study this, a Markov chain was generated from the transition probability function P_1 using 10,000 burn-ins. The chain following was one million trees long. The process took 6 hours on a workstation doing about 24 million floating point operations per second. The frequencies in our chain of the four topologies in Figure 5.1 from left to right were 56.53%, 36.51%, 2.42%, and 1.57%. There were

Figure 5.1: The experimental phylogeny using Bacteriophage T7. Node R constitutes the single outgroup taxon. The eight in-group taxa are denoted by the letters J–P. The topology in (a) is the true phylogeny. Topologies in (a)–(d) occur most frequently in the generated Markov chains.

31 other topologies generated in this chain, which together made up the remaining 2.97% of the trees. Our estimate of the mutation rate per site was 0.0043.

For diagnostic purposes we generated two chains which start at different initial trees. Our assessment of whether two chains converge to the same distribution is easy for this example. Because the distribution is dense in the four topologies in Figure 5.1 (making up about 97% of this distribution), we only need to compare the relative frequencies of those four topologies in both chains.

In this example the sequence *R* was used as an outgroup to root trees. We were only interested in the relationship among the eight in-group taxa. There were 56 possible rooted subtrees of size three. Table 5.1 shows the triples distances between any two topologies in Figure 5.1, which are proportions of triples of taxa whose relationships differ in the two topologies.

The first topology and the second topology in Figure 5.1 have minimum radius, which is .036, in order to cover at least 90% of trees in the chain. The 90% credible region is a subspace of trees in which $\max\{\rho(T, T_1), \rho(T, T_2)\} \leq .036$, where T is the

	(a)	(b)	(c)	(d)
(a)	—			
(b)	.036	—		
(c)	.036	.071	—	
(d)	.071	.036	.036	—

Table 5.1: The triples distances of the topologies in Figure 5.1.

topology of any tree in the chain and $T_i, i = 1, 2$ are the first and the second topology in Figure 5.1, respectively. However only the first topology has the least radius, which is still .036, in order to cover at least 95% of trees in the chain. The 95% credible region is a subspace of trees in which $\rho(T, T_1) \leq .036$, where T is the topology of any tree in the chain and T_1 is the first topology. The true topology is within both the 90% and the 95% credible regions. The true topology is also the modal tree and consensus tree in the chain.

We also generated a Markov chain of one million trees using the transition probability function P_2 . This chain worked more efficiently with the process taking 67 minutes on the same workstation. The frequencies that correspond to the four topologies in Figure 5.1 for this second method were 53.71%, 38.24%, 2.82%, and 2.04%. The estimate of the mutation rate per site was 0.0040. The 90% and 95% credible regions that are constructed based on the chain generated using P_2 are the same as ones that are constructed based on the chain generated using P_1 .

P_2 converges faster if no outgroup is prespecified than if an outgroup is prespecified. For example in the Bacteriophage T7 example P_2 converges after 10,000 burn-ins and 100,000 trees in the chain. It took less than 7 minutes to generate the chain. For diagnostic purposes we still generate two chains. For each chain the true topology was

used as the set A to create regeneration times. Figure 5.2 shows the SRQ plots for each chain. And Figure 5.3 shows the histograms of distances of 10,000 pairs of trees within and between two independent Markov chains. Neither of these two diagnostic tests shows any sign of divergence.

Figure 5.2: SRQ plots of T_i/T_n (vertical Axes) Against i/n (Horizontal Axes) for Two Markov Chains for the Bacteriophage data. Lines through the origin with unit slope are shown in both plots for comparisons.

To compare our results with other competitive methods, we created a consensus tree out of the chain generated using P_2 and created three consensus trees out of 500 bootstrap topologies using each of three phylogenetic construction methods: Neighbor-joining, Parsimony, and Maximum Likelihood (using programs in PHYLIP).

To make the comparisons more appropriate, we generated the neighbor-joining and maximum likelihood trees, using the Jukes-Cantor's one-parameter model. However,

Figure 5.3: Histograms of distances of 10,000 pairs of trees within and between two independent Markov chains for the Bacteriophage data. The two chains were generated using P_2 without assuming R as the outgroup.

the bootstrap results using Kimura's two-parameter model are close to the results using the one-parameter model in this case. Unrooted trees are generated using these methods. Since R is an outgroup taxon, the trees are rooted when a consensus tree is built. Figure 5.4 shows the consensus trees resulting from the four techniques.

The numbers listed at each fork in the tree give the proportions of simulated trees that duplicate the descendant subtree. The true tree was the consensus tree only for the Markov chain and the bootstrap parsimony methods with the part of the tree joining O and P being the least well specified for every technique. Our Markov chain Monte Carlo method is consistent (as $m \rightarrow \infty$) under the assumed model of evolution. This is not true of parsimony methods (see Felsenstein 1978b or Chang 1996). Although Maximum Likelihood provides a consistent estimate, finding a global maximum and carrying out a bootstrap analysis of the results is computationally prohibitive even for a moderate number of taxa.

Figure 5.4: Consensus trees for the Bacteriophage data using four tree construction methods.

5.2 The 16 species of fish data

In this example we analyze aligned 18S rRNA sequences of the 16 species of fish¹. Each sequence has 1789 sites. These 16 species of fish include three Cichlidae species (*Gaurochromis* sp., *Paretroplus polyactis*, and *Cichlasoma cyanoguttatum*), a marine Pomacentrid damselfish (*Pomacentrus melanochir*), two freshwater teleost taxa (*Etheostoma zonale* and a member of the teleost subfamily Acipenseformes, *Acipenser fulvescens*), one sea bass (*Sebastolobus altivelis*), one killifish (*Fundulus*

¹I am grateful to Dr. Paul Fuerst in the Department of Genetics at The Ohio State University for offering this data set.

heteoclitus), one coelacanth (*Latamaria chalumnae*), three sharks (*Notorynchus cepedianus*, *Squalus acanthia*, and *Echinorhinus cookei*), one ray (*Rhinobatos lentiginosus*), two lampreys (*Lampetra aepyptera* and *Petromyzon marinus*), and one tunicate (*Styela plicata*). Table 5.2 shows pairwise distances computed using Kimura's two-parameter model with a transition/transversion ratio of 2 (see page 51 of Li and Graur 1991). The distances were calculated using the DNA distance program in Felsenstein's PHYLIP. Among those 1789 sites, there are 76 sites at which there exists either a "-" or a "N" for some species. The "-" means that an insertion/deletion occurs during alignment. The "N" means unknown. In our implementation of the Markov chain Monte Carlo method, we ignored those sites. Without those sites, we calculate the proportions of pairwise differences among the sequences of the 16 species of fish. Table 5.3 shows the lower diagonal of the matrix.

For diagnostic purposes, we generated two Markov chains which start at different initial trees. For each chain, we generated 1.1 million trees using P_2 and discarded the first 0.1 million trees as a burn-in to ensure convergence. There were 84 and 75 distinct topologies in the first chain and the second chain, respectively. Our estimate of the mutation rate was 0.032 in both chains. The SRQ plots are shown in Figure 5.5. Here, we chose the set A to be the topology that occurred most frequently in the Markov chain in a primary run. And the histograms of the distances of trees within chains and between chains are shown in Figure 5.6. The SRQ plot showed no sign of divergence, while the inter-value distance histograms show a slight difference in the tree distributions. Perhaps a longer chain was needed in this case. The most frequent trees are shown in Figure 5.7. The frequencies of the topologies from the left to the right in Figure 5.7 estimated from both chains are given in Table 5.4.

	Sty	Eth	Pom	Par	Gau	Cic	Fun	Seb	Lat	Squ	Not	Rhi	Ech	Lam	Pet	Aci
Sty	—															
Eth	.1477	—														
Pom	.1532	.0092	—													
Par	.1510	.0121	.0115	—												
Gau	.1497	.0103	.0098	.0017	—											
Cic	.1490	.0116	.0092	.0029	.0011	—										
Fun	.1518	.0226	.0256	.0273	.0255	.0239	—									
Seb	.1505	.0185	.0244	.0232	.0226	.0233	.0350	—								
Lat	.1483	.0413	.0470	.0462	.0456	.0465	.0542	.0394	—							
Squ	.1433	.0365	.0414	.0395	.0389	.0397	.0486	.0262	.0292	—						
Not	.1441	.0334	.0384	.0365	.0359	.0366	.0468	.0250	.0268	.0074	—					
Rhi	.1443	.0480	.0499	.0491	.0486	.0494	.0591	.0405	.0388	.0291	.0267	—				
Ech	.1439	.0364	.0420	.0407	.0401	.0409	.0486	.0250	.0303	.0080	.0086	.0291	—			
Lam	.1527	.0853	.0869	.0879	.0867	.0858	.0899	.0912	.0873	.0828	.0809	.0834	.0827	—		
Pet	.1535	.0860	.0876	.0886	.0874	.0865	.0893	.0919	.0887	.0841	.0822	.0848	.0840	.0028	—	
Aci	.1680	.0556	.0582	.0593	.0575	.0590	.0664	.0664	.0683	.0658	.0633	.0772	.0651	.1070	.1064	—

Table 5.2: Shown on the lower diagonal are the corrected genetic distances using Kimura corrections for multiple substitutions with transition/transversion ratio of 2. Key: Eth= *Etheostoma*, Pom= *Pomacentrus melanochir*, Par= *Paretroplus*, Gau= *Gaurochromis* sp., Cic= *Cichlasoma cyanoguttatum*, Fun= *Fundulus heteoclitus*, Seb= *Sebastolobus altivelis*, Lat= *Latamaria chalumnae*, Squ= *Squalus acanthia*, Not= *Notorynchus cepedianus*, Rhi= *Rhinobatos lentiginosus*, Ech= *Echinorhinus cookei*, Lam= *Lampetra aepyptera*, Pet= *Petromyzon marinus*, Sty= *Styela*.

	Sty	Eth	Pom	Par	Gau	Cic	Fun	Seb	Lat	Squ	Not	Rhi	Ech	Lam	Pet	Aci
Sty	—															
Eth	.1290	—														
Pom	.1325	.0064	—													
Par	.1313	.0105	.0105	—												
Gau	.1308	.0088	.0088	.0018	—											
Cic	.1302	.0093	.0093	.0023	.0006	—										
Fun	.1331	.0181	.0204	.0228	.0210	.0204	—									
Seb	.1319	.0175	.0210	.0210	.0204	.0210	.0309	—								
Lat	.1302	.0368	.0403	.0403	.0397	.0403	.0485	.0356	—							
Squ	.1255	.0333	.0362	.0356	.0350	.0356	.0450	.0239	.0263	—						
Not	.1261	.0309	.0339	.0333	.0327	.0333	.0426	.0228	.0239	.0070	—					
Rhi	.1267	.0438	.0438	.0438	.0432	.0438	.0537	.0368	.0374	.0263	.0245	—				
Ech	.1261	.0333	.0368	.0368	.0362	.0368	.0450	.0228	.0274	.0082	.0082	.0263	—			
Lam	.1319	.0759	.0765	.0771	.0765	.0759	.0788	.0794	.0759	.0724	.0712	.0730	.0724	—		
Pet	.1331	.0771	.0776	.0782	.0776	.0771	.0788	.0806	.0771	.0736	.0724	.0741	.0736	.0023	—	
Aci	.1442	.0496	.0514	.0531	.0514	.0520	.0595	.0578	.0601	.0590	.0566	.0683	.0584	.0928	.0928	—

Table 5.3: Shown on the lower diagonal are the pairwise differences between any two sequences of the 16 species of fish. Key: Eth= Etheostoma, Pom= Pomacentrus melanochir, Par= Paretroplus, Gau= Gaurochromis sp., Cic= Cichlasoma cyanoguttatum, Fun= Fundulus heteoclitus, Seb= Sebastolobus altivelis, Lat= Latamaria chalumnnae, Squ= Squalus acanthia, Not= Notorynchus cepedianus, Rhi= Rhinobatos lentiginosus, Ech= Echinorhinus cookei, Lam= Lampetra aepyptera, Pet= Petromyzon marinus, Sty= Styela.

Figure 5.5: SRQ plots of T_i/T_n (vertical Axes) Against i/n (Horizontal Axes) for Two Markov Chains for the 16 species of fish data. Lines through the origin with unit slope are shown in both plots for comparisons.

Figure 5.6: Histograms of distances of 10,000 pairs of trees within and between two independent Markov chains for the 16 species of fish data.

	Topology			
Chain	(a)	(b)	(c)	(d)
1.	0.503	0.157	0.116	0.100
2.	0.486	0.148	0.107	0.098

Table 5.4: The frequencies of the four most frequent topologies (listed in Figure 5.7) occurred in the two chains.

The frequencies of the remaining topologies in the first chain together made up 12.4% of the total with each occurring less than 3% of the time. Similarly, the frequencies of the remaining topologies in the second chain together made up 16.1% with each being less than 3% except one topology which made up 6%.

To compare with competitive methods, we created 500 bootstrap trees using UPGMA, Neighbor-Joining and Maximum Parsimony methods. Each UPGMA tree, or Neighbor-Joining tree or Parsimony tree was created based on the two-parameter model with a transition/transversion ratio of 2 using the programs in PHYLIP. All four consensus trees shown in Figure 5.8 were created using the program consensus.c in PHYLIP. Among the four consensus trees, the Markov chain Monte Carlo consensus tree and the UPGMA consensus tree have the same topology except for the subtree involving the three species *Notorynchus*, *Squalus*, and *Echinorhinus*. The UPGMA consensus tree is the fourth most frequent topology occurred in the chain generated using the Markov chain Monte Carlo method. The Neighbor-Joining consensus tree and the Parsimony consensus tree have the same topology. Four consensus trees reveal two major clades. The five species, *Notorynchus*, *Echinorhinus*, *Squalus*, *Rhinobatos*, and *Latamaria*, are in one clade. The seven species, *Paretroplus*, *Gaurochromis*, *Cichlasoma*, *Pomacentrus*, *Etheostoma*, *Fundulus*, and *Sebastolobus*, are in the other

(a)

(b)

(c)

(d)

Figure 5.7: Topologies in (a)-(d) occur most frequently in the generated Markov chains for the aligned sequences of the 16 species of fish.

Markov chain Monte Carlo

UPGMA

Neighbor-Joining

Parsimony

Figure 5.8: Consensus trees for the 16 species of fish data using four tree construction methods.

clade. The species *Acipenser* is the outgroup of the two major clades in the Markov chain Monte Carlo and UPGMA consensus trees and belongs to the second clade in the Neighbor-joining and Parsimony consensus trees. The remaining three species, *Petromyzon*, *Lampetra*, and *Styela*, are the outgroups in all four consensus trees.

Which consensus tree is more likely to be the right answer? Examining the genetic distances using Kimura's two-parameter model with a transition/transversion ratio of 2 in Table 5.2 and the genetic distances without corrections in Table 5.3, we found that the distances between *Acipenser* and the species in the first and the second clades were greater than the distance between species within or between the two clades. So based on the one-parameter or the two-parameter model, *Acipenser* more likely appears to be an outgroup of the first and the second clades. The other disagreement between the topology of the Markov chain Monte Carlo/UPGMA consensus tree and the topology of the Neighbor-joining/Parsimony consensus tree is a subtree involved three species, *Pareuroplus*, *Gaurochromis*, and *Cichlasoma*, in the first clade. *Pareuroplus* and *Gaurochromis* were closer in the Neighbor-Joining/Parsimony consensus tree, while *Gaurochromis* and *Cichlasoma* were closer in the Markov chain Monte Carlo/UPGMA consensus tree. The distance between *Gaurochromis* and *Cichlasoma* is smaller than the distance between *Pareuroplus* and *Gaurochromis* either using Kimura corrections or without corrections of multiple substitutions. *Gaurochromis* is more likely closer to *Cichlasoma* than to *Pareuroplus*. Thus, the Markov chain Monte Carlo consensus tree or the UPGMA consensus tree describes a better relationship among the 16 species of fish given their observed sequences than the Neighbor-Joining consensus tree or the Parsimony consensus tree does.

5.3 The viral sequences from HIV infected Russian children

The data set used in our next example is from HIV-1 V3 sequences of HIV infected children in the former Soviet Union. During eleven months in 1988-1989 more than ninety children in Elista were infected through the use of contaminated needles from an HIV-infected infant. Some months later a similar outbreak of HIV-1 infection occurred in children in Rostov-on-Don (Russia), after some children from Elista had been transported to Rostov-on-Don hospitals. Bobkov et al. 1994 studied those incidences using nucleotide sequences from ten HIV-infected children in Elista, twelve of those infected in Rostov-on-Don, and eight unrelated HIV-seropositive individuals from Minsk. Each sequence consisted of 105 sites. Table 5.5 shows the aligned sequences excerpted from Bobkov et al. 1994. Another subject, RZM, was the mother of one of the infected children, RZD. RZM and RZD suggested a novel hypothesis: the transmission from infant to mother during breast feeding. The phylogenetic relationship suggested by the nucleotide sequence data could yield some information about this hypothesis. An RZD-to-RZM transition would suggest that they cluster together in the true phylogeny. However, the true phylogeny was unknown, so we used the Markov chain Monte Carlo method to estimate the probability distribution of the phylogeny.

The proportions of differences between Elista sequences in (a) and Rostov-on-Don sequences in (b) shown in Table 5.5 were only 0 – 10%. The proportions of difference between a sequence in (a) or (b) and a Minsk sequence in (c) were about 11 – 26%. The proportions of differences between sequences within (c) were in the range 1 – 25%. There were five subjects from Elista and Rostov-on-Don whose sequences were identical to the sequence of the subject RZM. We excluded those five sequences and

		680		700		720		740		760		
		.		.		.		.		.		
(a)	R/E	TGTACCAGACCCAATAACAATACAAGAAAAAGTATAAGTTTAGGACCAGGACAAGCGTTCTATACAACAGGTGAATAATAGGAGATATAAGACAAGCACATTGT										62
	RZM	-----										6
	BLS	-----C-----										2
	ZHR	-----										8
	KLB	-----A-----GG--G-----										8
	ODN	-----T-----G--G-----C-T-----G-----										8
	SHS	-----										6
	ALV	-----C-----G-----										8
	PCH	-----										9
	RZD	-----T-----C--G-----										4
	KRV	-----CA-G-T-----A--G-----										3
(b)	R/E	TGTACCAGACCCAATAACAATACAAGAAAAAGTATAAGTTTAGGACCAGGACAAGCGTTCTATACAACAGGTGAATAATAGGAGATATAAGACAAGCACATTGT										65
	PVL	-----A-----										4
	UMD	-----A-----G-----G-----C-----C-----										5
	VRN	-----										5
	IZD	-----G-----										5
	TRK	-----										3
	KHL	-----A-----C-----C-----G-----										9
	SRN	-----A-T-----A-----C-----										6
	SND	-----										4
	BDZ	-----T-----C-----C-----										6
	OVC	-----G-----C-----T-----										5
	MCH	-----GA-----										6
	LNT	-----C-C-----G--T-----										7
(c)	NC	TGTACAAGACCCAACAACAATACAAGAAAAAGTATACATATAGGACCAGGAGAGCATTTCTATGCAACAGGAGACATAATAGGAGATATAAGACAAGCAGATTGT										38
	ZLN	-----G-----T-----G--C--AGG-----ACA-A-----C-----										5
	KZL	-----T-----T-----T-----A-----C-----										4
	KTS	-----G-----T-----C--A-G-----ACA-A-----C-----AA-----										6
	KLT	-----C-----T-----C-----GC-----ACA-T-----T-G-----A-GA-----A--TT-----										5
	KTC	-----CT--G--T-----C--G--C--G-----C--C--T-A-----C-----A-----										4
	CHS	-----										6
	SRC	-----C-----G-----A--GG--A-----										5
	LPK	-----G-----T-----T-----A-----T-----										5
(d)	R/E	TGTACCAGACCCAATAACAATACAAGAAAAAGTATAAGTTTAGGACCAGGACAAGCATTTCTATACAACAGGTGAATAATAGGAGATATAAGACAAGCACATTGT										
	CONRZM	-----										
	207RZM	-----										
	208RZM	-----G-----										
	194RZM	-----G-----										
	196RZM	-----										
	198RZM	-----										
	210RZM	-----G-----										
(e)	CONRZD	-----T-----C--G-----										
	267RZD	-----										
	269RZD	-----T-----C--G-----										
	278RZD	-----A--T-----AA--G--T-----										
	280RZD	-----T-----CA--C-----C--G-----										

Table 5.5: Aligned sequences in the V3 region. The sequences from (a) Elista (n=10)(b) Rostov-on-Don (n=12) and (c) Minsk (n=8) patients (d) the sequences of a mother (RZM) and (e) the sequences of her HIV-infected daughter (RZD). The nucleotides identical to the consensus sequences are indicated by a dash; and those which are different from the consensus are represented by letters. The three letters on the left column represent the identification code for each individual. The number on the far right column represents the number of clones which have been sequenced for each subject.

	BLS	KLB	ODN	ALV	RZD	KRV	PVL	UMD	IZD	KHL	SRN	SND	BDZ	OVC	MCH	LNT	ZLN
RZM	.010																
BLS	.038	.048															
KLB	.057	.067	.095														
ODN	.019	.029	.057	.076													
ALV	.029	.038	.048	.086	.038												
RZD	.057	.067	.048	.114	.067	.048											
KRV	.010	.019	.048	.067	.029	.038	.057										
PVL	.038	.048	.057	.095	.057	.067	.076	.048									
UMD	.010	.019	.048	.067	.029	.038	.067	.019	.048								
IZD	.038	.029	.057	.095	.038	.048	.076	.048	.057	.048							
KHL	.038	.029	.057	.095	.048	.038	.057	.048	.057	.048	.029						
SRN	.010	.019	.048	.067	.029	.038	.067	.019	.048	.019	.048	.048					
SND	.029	.019	.067	.086	.029	.048	.076	.038	.067	.038	.048	.038	.038				
BDZ	.029	.019	.067	.086	.048	.057	.086	.038	.067	.038	.048	.048	.019	.038			
OVC	.019	.029	.057	.076	.038	.048	.067	.010	.057	.029	.057	.057	.029	.048	.029		
MCH	.038	.048	.076	.095	.019	.057	.086	.048	.076	.048	.057	.067	.029	.048	.048	.057	
LNT	.124	.114	.143	.171	.143	.133	.152	.114	.162	.133	.143	.143	.133	.124	.114	.105	.162

Table 5.6: Proportions of pairwise differences among 18 HIV-1 V3 sequences from HIV infected Russian children.

used the sequence RZM and the rest of the sequences from Elista and Rostov-on-Don to estimate the probability that the sequence RZM was closer to the sequence RZD than to the other sequences. And we also selected the sequence ZLN from Minsk as the outgroup to root the tree. In total, 18 sequences were used in our analysis. The proportions of pairwise differences among those sequences are shown in Table 5.6.

Two chains were generated using P_2 . During the generating process, we did not prespecify an outgroup, so we could also test whether the sequence from Minsk was truly the outgroup to the sequences from Elista and Rostov-on-Don. Our first running of P_2 in this example showed that the distribution of topologies in this example was much less dense than in the first two examples. For example, there were 32,484 distinct topologies out of 150,000 generated trees. Since the memory space required to save this many topologies almost reached the capacity of the workstation that we used, we decided to generate chains with 150,000 trees following a long burn-in of, say, 1 million trees.

In order to use the SRQ plot as a diagnostic, we needed to choose an appropriate set A to generate appropriate numbers of regeneration times. In the cases where the distribution of trees is so dispersed, as in this example, any single topology forms too small a set to create enough regeneration times. In this case we chose a set of many topologies as the set A . In both chains we chose the last burn-in tree as the center of the set. The set A in the first chain included the trees that differed from the center tree by no more than 500 rooted subtrees of size three, or within a triples distance of 0.74. The set A in the second chain included the trees that differed from the center tree by no more than 450 rooted subtrees of size three, or within a triples distance of 0.62. With the set A defined in this way, we got 5610 regeneration times in the first

chain and 6294 in the second chain. The two SRQ plots are shown in Figure 5.9. The SRQ plots are apparently not close to the straight lines. The flat parts of SRQ curves indicate that the set A is too large. Thus those SRQ plots may not be appropriate as diagnostic tests of convergence. However we can always use the inter-value distance diagnostic. The three histograms of distances of the trees between the two chains and the trees within the chains in Figure 5.10 show no sign of divergence.

Figure 5.9: SRQ plots of T_i/T_n (vertical Axes) Against i/n (Horizontal Axes) for Two Markov Chains for the viral sequences data collected from the HIV infected Russian children. Lines through the origin with unit slope are shown in both plots for comparisons.

There were 33033 distinct topologies in the first chain and 32637 distinct topologies in the second chain. 99% of the trees in the first chain and 97% of the trees in the second chain showed that the sequence ZLN from Minsk was the outgroup. This

Figure 5.10: Histograms of distances of 10,000 pairs of trees within and between two independent Markov chains. The chains were generated using P_2 for the viral sequences data collected from the HIV infected Russian children.

suggested that the source of infections of the children in Elista and Rostov-on-Don was not from Minsk (or at least from a sequence closely related to ZLN). The phylogenetic relationship among the seventeen sequences from Elista and Rostov-on-Don was not apparent. Every topology occurred either in the first chain or in the second chain no more than 0.4% of the time. In the first chain the forty-two most frequent topologies together accounted for only 1% of the distribution; the top 6716 most frequent topologies made up 50% of the empirical distribution, and the top 21765 made up 90%. In the second chain the forty two most frequent topologies together accounted for 1%; the top 6642 most frequent topologies made up 50% of the empirical distribution, and the top 21467 made up 90%. The cumulative frequencies of the topologies from the most frequent to the least frequent in both chains were shown in Figure 5.11.

Was the topology of the 17 sequences from Elista and Rostov-on-Don distributed

Figure 5.11: Two cumulative frequencies of the topologies from the most frequent to the last frequent in the first chain and the second chain.

randomly? According to Critchlow, Pearl and Qian 1996 the triples distance of trees of size 17 under the uniform model (i.e. when the underlying probability distribution of topologies is uniform) is approximately normally distributed for large tree size with approximate mean $\frac{2}{3}$ and approximate standard deviation 0.070. The estimated mean and standard deviation of triples distances between topologies were .549 and .127 from the first chain and .549 and .132 from the second chain. These results indicated that the topologies were almostly randomly distributed.

Assuming that the sequence ZLN from Minsk is the outgroup, next we test whether the RZM and RZD pair are closer to each other in the true phylogeny than to the other sequences from either Elista or Rostov-on-Don. If RZM is randomly clustered with any other 16 sequences in the true phylogeny, the probability that RZM is clustered

with RZD is $\frac{\pi(2n-5)}{\pi(2n-3)}$ (about 3.2%), where n is the number of in-group taxa, i.e. 17 in this case. However the probability that RZM and RZD clustered in the true phylogeny was 0.91% estimated based on the first chain or 0.98% estimated based on the second chain. These results suggest that we should not accept the hypothesis of transmission from infant to mother through breast feeding. However, one of the minor strains in the intra-subject sequence population of RZD (see Table 5.5(e)) was identical to the sequence of RZM used in our analysis. This fact perhaps suggests a transmission between RZM and RZD though in an unknown direction. In summary, the results using the Markov chain Monte Carlo method do not contradict the conclusion stated in Bobkov et al. (1994) that

“These data are compatible with the postulated route of HIV-1 transmission from infant to mother during breast feeding, although infection of the mother by another route, such as parenteral transmission cannot be completely excluded.”

CHAPTER 6

DISCUSSION AND EXTENSIONS

In chapter 5 we applied the Markov chain Monte Carlo method to three data sets. For the first two data sets, we created the consensus tree from the chain and compared it with the Neighbor-joining consensus tree, the Parsimony consensus tree, UPGMA consensus tree, and Maximum Likelihood consensus tree. The Markov chain Monte Carlo method produced a more reliable consensus tree than the other methods, because the chain generated using the Markov chain Monte Carlo converges to the conditional probability distribution of trees given the observed nucleotide sequences of the taxa. The consensus trees based on the other methods require bootstrapping. For large numbers of taxa, the Parsimony method and the Maximum Likelihood method rely on an incomplete heuristic and on time consuming search strategies that do not guarantee producing an absolute maximum.

In summary, a key advantage of the Markov chain simulation technique is the direct and relevant probabilistic interpretation of the results. In particular the method estimates the probability that the simulated tree is the true tree for these sequences. The distribution of trees generated by the Markov chain can be directly translated to describe the randomness in any aspect of the phylogenetic relationship. For example, the credible region of the true phylogeny in Section 4.4 and the reliability test of

subtrees in Section 4.3. Estimation of the mutation rate per site is a by-product of the Markov chain Monte Carlo method. Thus, our techniques should be a valuable tool for the molecular geneticist seeking to quantify the degree of uncertainty in an estimated phylogeny.

Our algorithm, in the current implementation, assumes an equal mutation rate. In many applications, this assumption is not tenable. It may be important to extend the algorithms in their implementation to incorporate a more general model of evolution. For example, allowing for different substitution rates for each combination of nucleotides gives the 12-parameter model discussed in Rodriguez et al. (1990). The estimation of the 12 substitution rates needed for the modified formula can be accomplished efficiently during the burn-in period using estimates analogous to (3.17).

The 12-parameter model is generally considered an overfit to most molecular data (Rodriguez et al. 1990). The 5-parameter model of Hasegawa Yano, and Kishino (1984), which allows for differing rates of mutation for transitions and transversions as well as differing equilibrium probabilities for each nucleotide is a good compromise. This was the model assumed in one application of the algorithm of Mau, Newton and Larget (1996). (Their algorithm used a Bayesian approach. The chain generated using their algorithm is interpreted as converging to the posterior distribution of trees.) To extend the implementation of our algorithm from the one-parameter to five-parameter model, we only need to modify the formula for the summation in Equation 2.4 to incorporate the transition probability matrix given in 2.10.

The Markov chain Monte Carlo method described in this dissertation has several other possible extensions in the future. For example, in our probability model we made the assumption of a molecular clock while some other competitive methods do

not. To allow the mutation rate to vary in different branches, we need to modify the probability distribution of trees accordingly. Given the probability distribution of trees, the rest of the algorithm can easily be adjusted accordingly. We can also modify our distribution of trees to allow the mutation rate to vary over sites following the model in Yang (1993), (1994), and (1995) or following Felsenstein and Churchill (1996) who allow a variable mutation rate over sites in the Maximum Likelihood method. Future extensions of our method might also include relaxing the assumption of independence between sites and applications based on amino acid sequences as well.

For all these possible extensions, the form of the probability distribution becomes complicated, but the calculation of the transition probabilities remains tractable. We should be able to run the chain in a reasonable amount of time. This is true especially for P_2 .

BIBLIOGRAPHY

- [1] K. Athreya, H. Doss, and J. Sethuraman. On the convergence of the Markov chain simulation method. *The Annals of Statistics*, 24:69–100, 1996.
- [2] A. Bobkov, M. Garaev, A. Rzhainova, P. Kaleebu, R. Pitman, J. N. Weber, and R. Cheingsong-Popov. Molecular epidemiology of HIV-1 in the former soviet union: analysis of ENV V3 sequences and their correlation with epidemiological data. *AIDS*, 8:619–624, 1994.
- [3] J. T. Chang. Inconsistency of evolutionary tree topology reconstruction methods when substitution rates vary across characters. *Mathematical Biosciences*, 134:189–215, 1996.
- [4] M. K. Cowles and B. P. Carlin. Markov chain Monte Carlo convergence diagnostics: A comparative review. Technical report, Department of Biostatistics, Harvard School of Public Health, 1995.
- [5] J. Cracraft and K. Helm-Bychowski. Parsimony and phylogenetic inference using DNA sequences: Some methodological strategies. In M. M. Miyamoto and J. Cracraft, editors, *Phylogenetic Analysis of DNA Sequence*. Oxford University Press, 1991.
- [6] D. E. Critchlow, D. K. Pearl, and C. Qian. The triples distance for rooted bifurcating phylogenetic trees. *Systematic Biology*, 45(3):323–334, 1996.
- [7] J. L. Doob. *Stochastic Processes*. Wiley, New York, 1953.
- [8] H. Doss, F. Huffer, and K. Lawson. Bayesian nonparametric estimation via Gibbs sampling for coherent systems with redundancy. Technical Report 894, Department of Statistics, Florida State University, 1994.
- [9] B. Efron. *The Jackknife, the Bootstrap and Other Resampling Plans*. Society for Industrial and Applied Mathematics, Philadelphia, Philadelphia, Pennsylvania, 1982.

- [10] B. Efron, E. Halloran, and S. Holmes. Bootstrap confidence levels for phylogenetic trees. Technical Report 489, Department of Statistics, Stanford University, 1995.
- [11] J. Felsenstein. Cases in which parsimony and compatibility methods will be positively misleading. *Systematic Zoology*, 27:401–410, 1978.
- [12] J. Felsenstein. The number of evolutionary trees. *Systematic Zoology*, 27:27–33, 1978.
- [13] J. Felsenstein. Evolutionary trees from DNA sequences: A maximum likelihood approach. *Journal of Molecular Evolution*, 17:368–376, 1981.
- [14] J. Felsenstein. Confidence limits on phylogenies: An approach using the bootstrap. *Evolution*, 39(4):783–791, 1985.
- [15] J. Felsenstein. *Phylogeny Inference Package (PHYLIP)*. University of Washington, Seattle, 3.5 edition, 1993.
- [16] J. Felsenstein. A hidden Markov model approach to variation among sites in rate of evolution. *Molecular Biology Evolution*, 13(1):93–104, 1996.
- [17] W. M. Fitch. Toward defining the course of evolution: Minimum change for a specific tree topology. *Systematic Zoology*, 20:406–416, 1971.
- [18] W. M. Fitch and E. Margoliash. Construction of phylogenetic trees. *Science*, 155:279–284, 1967.
- [19] J. Friedman and L. Rafsky. Graphics for the multivariate two-sample problem. *Journal of the American Statistical Association*, 76:277–287, 1984.
- [20] M. Hasegawa, H. Kishino, and T. Yano. Dating of the human-ape splitting by a molecular clock of mitochondrial DNA. *Journal of Molecular Evolution*, 22:160–174, 1985.
- [21] M. Hasegawa, T. Yano, and H. Kishino. A new molecular clock of mitochondrial DNA and the evolution of hominoids. *Proceedings of the Japanese Academy of Sciences, B*, 60:95–98, 1984.
- [22] D. M. Hillis, M. W. Allard, and M. M. Miyamoto. Analysis of DNA sequence data: Phylogenetic inference. *Methods in Enzymology*, 224:457–487, 1993.
- [23] D. M. Hillis and J. J. Bull. An empirical test of bootstrapping as a method for assessing confidence in phylogenetic analysis. *Systematic Biology*, 42:182–192, 1993.

- [24] D. M. Hillis, J. J. Bull, M. E. White, M. R. Badgett, and I. J. Molineux. Experimental phylogenetics: Generation of a known phylogeny. *Science*, 255:589–592, 1992.
- [25] D. M. Hillis, J. P. Huelsenbeck, and C. W. Cunningham. Application and accuracy of molecular phylogenies. *Science*, 264:671–677, 1994.
- [26] D. M. Hillis, B. Mable, and C. Moritz. Applications of molecular systematics: The state of the field and a look to the future. In D. M. Hillis, C. Moritz, and B. Mable, editors, *Molecular Systematics*, chapter 12, pages 515–543. Sunderland, Massachusetts Sinauer, second edition, 1996.
- [27] T. Jukes and C. Cantor. Evolution of protein molecules. In H. N. Munro, editor, *Mammalian Protein Metabolism*, pages 21–132. New York: Academic Press, 1969.
- [28] J. Kiefer. k -sample analogues of the Kolmogorov-Smirnov and Cramér-von Mises tests. *Annals of Mathematical Statistic*, 30:420–447, 1959.
- [29] M. Kimura. A simple method for estimating evolutionary rates of base substitutions through comparative studies of nucleotide sequences. *Journal of Molecular Evolution*, 16:111–120, 1980.
- [30] H. Kishino, T. Miyata, and M. Hasegawa. Maximum likelihood inference of protein phylogeny and the origin of chloroplasts. *Journal of Molecular Evolution*, 31:151–160, 1990.
- [31] M. K. Kuhner and J. Felsenstein. A simulation comparison of phylogeny algorithms under equal and unequal evolutionary rates. *Molecular Biology Evolution*, 11(3):459–468, 1994.
- [32] M. K. Kuhner, J. Yamato, and J. Felsenstein. Estimating effective population size and mutation rate from sequence data using Metropolis-Hastings sampling. *Genetics*, 140:1421–1430, 1995.
- [33] J. Lake. A rate-independent technique for analysis of nucleic acid sequences: Evolutionary parsimony. *Molecular Biology and Evolution*, 4:167–191, 1987.
- [34] W. H. Li and D. Graur. *Fundamentals of Molecular Evolution*. Sinauer Associates, Inc., Sunderland, Massachusetts, 1991.
- [35] J. F. Maa, D. K. Pearl, and R. Bartoszynski. Reducing multidimensional two-sample data to one-dimensional interpoint comparisons. *The Annals of Statistics*, 24(3):1069–1074, 1996.

- [36] T. Margush and F. R. McMorris. Consensus n -trees. *Bulletin Mathematics Biology*, 43:239–244, 1981.
- [37] B. Mau, M. A. Newton, and B. Larget. Bayesian phylogenetic inference via Markov chain Monte Carlo methods. Technical Report 961, Department of Statistics, University of Wisconsin-Madison, 1996.
- [38] P. Mykland, L. Tierney, and B. Yu. Regeneration in Markov chain samplers. *Journal of the American Statistical Association*, 90:233–241, 1995.
- [39] M. Nei. Relative efficiencies of different tree-making methods for molecular data. In M. M. Miyamoto and J. Cracraft, editors, *Phylogenetic Analysis of DNA Sequences*. Oxford University Press, 1991.
- [40] E. Nummelin. *General Irreducible Markov Chains and Non-Negative Operators*. Cambridge Univ. Press, 1984.
- [41] W. H. Press, S. A. Teukolsky, W. T. Vetterling, and B. P. Flannery. *Numerical Recipes in C: The Art of Scientific Computing*. Cambridge University Press, second edition, 1992.
- [42] F. Rodriguez, J. L. Oliver, A. Marin, and J. R. Medina. The general stochastic model of nucleotide substitution. *Journal of Theoretical Biology*, 4:406–425, 1990.
- [43] N. Saitou and M. Nei. The Neighbor-joining method: A new method for reconstructing phylogenetic trees. *Molecular Biology Evolution*, 4:406–425, 1987.
- [44] R. T. Schuh and J. T. Polhemus. Analysis of taxonomic congruence among morphological, ecological, and biogeographic data sets for the leptocondomorpha (hemiptera). *Systematic Zoology*, 29:1–26, 1981.
- [45] R. R. Sokal and C. Michener. A statistical method for evaluating systematic relationships. *Univ. Kansas Sci. Bull.*, 28:1409–1438, 1958.
- [46] R. R. Sokal and F. J. Rohlf. Taxonomic congruence in the leptocondomorpha reexamined. *Systemic Zoology*, 36:175–195, 1981.
- [47] D. L. Swofford, G. J. Olsen, P. J. Waddell, and D. M. Hillis. Phylogenetic inference. In D. M. Hillis, C. Moritz, and B. Mable, editors, *Molecular Systematics*, chapter 11, pages 407–514. Sunderland, Massachusetts Sinauer, second edition, 1996.
- [48] Y. Tateno, M. Nei, and F. Tajima. Accuracy of estimated phylogenetic trees from molecular data. *Journal of Molecular Evolution*, 18:387–404, 1982.

- [49] L. Tierney. Markov chains for exploring posterior distributions. *The Annals of Statistics*, 22:1701–1727, 1994.
- [50] Z. Yang. Maximum likelihood estimation of phylogeny from DNA sequences when substitution rates differ over sites. *Molecular Biology Evolution*, 10:1396–1401, 1993.
- [51] Z. Yang. Maximum likelihood phylogenetic estimation from DNA sequences with variable rates over sites: approximate methods. *Journal of Molecular Evolution*, 39:306–314, 1994.
- [52] Z. Yang. A space-time process model for the evolution of DNA sequences. *Genetics*, 139:993–1005, 1995.
- [53] A. Zharkikh and W. H. Li. Statistical properties of bootstrap estimation of phylogenetic variability from nucleotide sequences, I: four taxa with a molecular clock. *Molecular Biology Evolution*, 9:1119–1147, 1992.
- [54] A. Zharkikh and W. H. Li. Statistical properties of bootstrap estimation of phylogenetic variability from nucleotide sequences, II: four taxa without a molecular clock. *Journal of Molecular Evolution*, 35:356–366, 1992.