Maximum Likelihood Estimation of Phylogenetic Tree with Evolutionary Parameters

DISSERTATION

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

By

Qiang Wang, B.S., M.S.

* * * * *

The Ohio State University

2004

Dissertation Committee:

Professor Dennis K. Pearl, Adviser
Professor Douglas E. Critchlow
Professor Hani Doss

Approved by

Adviser
Department of Statistics
ABSTRACT

This thesis deals with issues in the maximum likelihood estimation of phylogenetic tree construction. The maximum likelihood estimator (MLE) of evolutionary parameters is shown to be asymptotically efficient with the number of sites approaching infinity, when the tree is assumed to be known. Through simulation, a Monte Carlo Sampling method for approximating the expected Fisher information is demonstrated to be superior to using the observed Fisher information implemented in a phylogenetic analysis package to estimate the variability of MLE. Simulations also revealed the underestimation of the variability when the tree is unknown and needs to be simultaneously estimated. The large-sample property of the joint MLE for tree topology, branch lengths vector and evolutionary parameters is established. As the number of sites approaches infinity, the probability of recovering the true tree topology tends to one and the joint MLE for the branch lengths vector and evolutionary parameters are shown to be efficient. In order to estimate the variability of the MLE for evolutionary parameters, a standard bootstrap algorithm and a bootstrap method motivated by the conditional variance principle (CVB) are proposed for practical implementation. Based on simulation evidence, a parametric version of the bootstrap methods is recommended, and CVB method in particular performs reasonably well and provides substantial computational time savings. Lastly, the method is applied to nucleotide data of 30 papillomavirus strains.
In memory of my father
ACKNOWLEDGMENTS

I would like to thank my adviser, Professor Pearl, for his continued advice, encouragement and guidance throughout my thesis work and his patience and understanding for my particular circumstances. I also thank my committee members, Professor Critchlow and Professor Doss, for taking time to serve on my dissertation committee. I am grateful to Dr. Laura Salter for her collaboration and providing her SSA software. I wish to thank my manager at Abbott Labs, Dr. Charles Locke, for his support while I finish up the dissertation. Finally, I am extremely grateful to my wife for her sacrifices and support over the difficult times.
VITA

August 4, 1973 ........................... Born - Qingdao, China

1992 ................................. B.S. Applied Mathematics, Tsinghua University, Beijing, China


1995-1999 ............................. Graduate Research Associate, The Ohio State University.

2000-present .......................... Statistician, Abbott Laboratories.

PUBLICATIONS

Research Publications


FIELDS OF STUDY

Major Field: Biostatistics
# TABLE OF CONTENTS

<table>
<thead>
<tr>
<th>Section</th>
<th>Page</th>
</tr>
</thead>
<tbody>
<tr>
<td>Abstract</td>
<td>ii</td>
</tr>
<tr>
<td>Dedication</td>
<td>iii</td>
</tr>
<tr>
<td>Acknowledgments</td>
<td>iv</td>
</tr>
<tr>
<td>Vita</td>
<td>v</td>
</tr>
<tr>
<td>List of Tables</td>
<td>viii</td>
</tr>
<tr>
<td>List of Figures</td>
<td>x</td>
</tr>
<tr>
<td>Chapters:</td>
<td></td>
</tr>
<tr>
<td>1. Introduction</td>
<td>1</td>
</tr>
<tr>
<td>1.1 DNA and Nucleotide Sequences</td>
<td>2</td>
</tr>
<tr>
<td>1.2 Phylogenetic Tree</td>
<td>4</td>
</tr>
<tr>
<td>1.3 Phylogenetic Tree Reconstruction</td>
<td>7</td>
</tr>
<tr>
<td>1.4 Stochastic Models of Nucleotide Evolution</td>
<td>17</td>
</tr>
<tr>
<td>1.5 Maximum Likelihood Estimation</td>
<td>23</td>
</tr>
<tr>
<td>1.6 Evolutionary Parameters Estimation</td>
<td>27</td>
</tr>
<tr>
<td>2. Asymptotic Efficiency of Evolutionary Parameters when Tree is Known</td>
<td>32</td>
</tr>
<tr>
<td>2.1 Background</td>
<td>33</td>
</tr>
<tr>
<td>2.1.1 Notation</td>
<td>33</td>
</tr>
<tr>
<td>2.1.2 Consistency of ML Tree</td>
<td>34</td>
</tr>
<tr>
<td>2.2 Asymptotic Efficiency of MLE of Evolutionary Parameters</td>
<td>38</td>
</tr>
<tr>
<td>2.2.1 Birch’s Conditions</td>
<td>39</td>
</tr>
</tbody>
</table>
2.2.2 Theoretical Efficiency Results of MLE for Evolutionary Parameters .................................................. 41
2.3 Fisher Information ................................................................. 66
2.4 Simulation ......................................................................... 69
  2.4.1 Simulation 1 ................................................................. 71
  2.4.2 Simulation 2 ................................................................. 74
  2.4.3 Simulation 3 ................................................................. 82

3. Limiting Distribution of MLE when Tree is Unknown .................. 89
  3.1 Limiting Distribution of Joint MLE when the Tree Topology is Known 90
  3.2 Limiting Distribution of Joint MLE when the Tree Topology is Estimated ........................................ 104

4. Bootstrap Estimators of Variance for the Model Parameter MLE ..... 108
  4.1 Direct Bootstrap ............................................................... 111
  4.2 Conditional Variance Bootstrap ........................................ 112
  4.3 Simulation .................................................................. 115
  4.4 Application to HPV Sequences ....................................... 121
  4.5 Discussion .................................................................. 128

5. Future Directions ................................................................. 135
  5.1 Tree Construction Related Problems .................................. 135
  5.2 Simultaneous Alignment and Phylogeny Construction ........... 137
    5.2.1 Background ............................................................ 138
    5.2.2 A Proposal for Simultaneous Tree Inference and Sequence Alignment ............................................ 141
    5.2.3 Results ................................................................. 149
    5.2.4 Discussion ............................................................ 153

Bibliography ................................................................. 157
# LIST OF TABLES

<table>
<thead>
<tr>
<th>Table</th>
<th>Page</th>
</tr>
</thead>
<tbody>
<tr>
<td>1.1 DNA sequences for a portion of the L1 gene for seven Group A9 papillomaviruses.</td>
<td>4</td>
</tr>
<tr>
<td>2.1 Summary of the results for s.e. of $\mu$ in Simulations 1 and 2</td>
<td>80</td>
</tr>
<tr>
<td>2.2 Summary of the results for s.e. of $K$ in Simulations 1 and 2</td>
<td>80</td>
</tr>
<tr>
<td>2.3 Summary of the results for s.e. of $\mu$ in Simulations 2 and 3</td>
<td>87</td>
</tr>
<tr>
<td>2.4 Summary of the results for s.e. of $K$ in Simulations 2 and 3</td>
<td>88</td>
</tr>
<tr>
<td>4.1 Summary of results for nonparametric bootstrap methods for $\mu$</td>
<td>117</td>
</tr>
<tr>
<td>4.2 Summary of results for nonparametric bootstrap methods for $K$</td>
<td>117</td>
</tr>
<tr>
<td>4.3 Summary of results for parametric bootstrap methods for $\mu$</td>
<td>117</td>
</tr>
<tr>
<td>4.4 Summary of results for parametric bootstrap methods for $K$</td>
<td>118</td>
</tr>
<tr>
<td>4.5 Summary of pseudo-MSE for bootstrap s.e. estimator for $\mu$</td>
<td>119</td>
</tr>
<tr>
<td>4.6 Summary of pseudo-MSE for bootstrap s.e. estimator for $K$</td>
<td>119</td>
</tr>
<tr>
<td>4.7 MLE of $\mu$ and $K$ and CVB estimates of their s.e.’s for the papillomavirus data set</td>
<td>125</td>
</tr>
<tr>
<td>5.1 Example for three hypothetical unaligned nucleotide sequences</td>
<td>137</td>
</tr>
</tbody>
</table>
5.2 Summary of the expected vs. observed cell counts for the example with analytical solution . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . 150
# LIST OF FIGURES

<table>
<thead>
<tr>
<th>Figure</th>
<th>Description</th>
<th>Page</th>
</tr>
</thead>
<tbody>
<tr>
<td>1.1</td>
<td>Unrooted and rooted phylogenetic trees with five OTUs</td>
<td>6</td>
</tr>
<tr>
<td>1.2</td>
<td>The local rearrangement transition strategy</td>
<td>13</td>
</tr>
<tr>
<td>1.3</td>
<td>A rooted tree with four external nodes</td>
<td>24</td>
</tr>
<tr>
<td>2.1</td>
<td>Illustration of an ((n + 1))-external-node rooted tree and the tree with an external node removed</td>
<td>58</td>
</tr>
<tr>
<td>2.2</td>
<td>ML tree for the 14-sequence mtDNA data</td>
<td>70</td>
</tr>
<tr>
<td>2.3</td>
<td>Results of Simulation 1</td>
<td>73</td>
</tr>
<tr>
<td>2.4</td>
<td>Results of Simulation 2, with (\mu = 0.1) and (K = 0.5)</td>
<td>76</td>
</tr>
<tr>
<td>2.5</td>
<td>Results of Simulation 2, with (\mu = 0.1) and (K = 10)</td>
<td>77</td>
</tr>
<tr>
<td>2.6</td>
<td>Results of Simulation 2, with (\mu = 0.5) and (K = 0.5)</td>
<td>78</td>
</tr>
<tr>
<td>2.7</td>
<td>Results of Simulation 2, with (\mu = 0.5) and (K = 10)</td>
<td>79</td>
</tr>
<tr>
<td>2.8</td>
<td>Results of Simulation 3, with (\mu = 0.1) and (K = 0.5)</td>
<td>83</td>
</tr>
<tr>
<td>2.9</td>
<td>Results of Simulation 3 with (\mu = 0.1) and (K = 10)</td>
<td>84</td>
</tr>
<tr>
<td>2.10</td>
<td>Results of Simulation 3, with (\mu = 0.5) and (K = 0.5)</td>
<td>85</td>
</tr>
<tr>
<td>2.11</td>
<td>Results of Simulation 3, with (\mu = 0.5) and (K = 10)</td>
<td>86</td>
</tr>
</tbody>
</table>
3.1 Illustration of an $N$-external-node unrooted tree and the tree with an external node removed ........................................... 93

3.2 Illustration of a three-external-node unrooted tree and the tree with an external node removed ........................................... 95

4.1 ML trees for papillomavirus data set ........................................... 124

4.2 Other high-likelihood trees for papillomavirus data set ................. 127

5.1 Three possible alignments with phylogenies for the three hypothetical sequences ......................................................... 138

5.2 Illustration of drawing a random pairwise alignment .................... 147
CHAPTER 1

INTRODUCTION

Study of phylogeny, the evolutionary relationship among the organisms, can be dated back over 200 years ago when biologists built trees to classify species based on morphological data. Since the late 1950s, various molecular sequencing techniques have been developed, resulting in the rapid accumulation of DNA sequence data, especially after the invention of the Polymerase Chain Reaction (PCR) technique by Mullis in 1985. Consequently, phylogenetic analysis based on molecular data has become increasingly popular. The results of phylogenetic analyses are useful in studying the evolution of species, individual relatedness, geographic variation, as well as in understanding viral transmission and treatment resistance development. The enhancement of modern computers allows computationally intensive tree-building methods, most notably the maximum likelihood method, to be practically implemented and widely used. In this thesis, we will prove asymptotic efficiency results for reconstructing phylogenetic trees and estimating evolutionary parameters via the maximum likelihood approach, and also study a practical method to estimate the variability associated with the MLE of evolutionary parameters.

In this Chapter, the basic concepts and notations for nucleotide sequences, phylogenetic tree and various schools of tree building will be reviewed.
1.1 DNA and Nucleotide Sequences

Deoxyribonucleic acid (DNA) molecules carry the hereditary information for nearly all living organisms. DNA is a double helix polymer with two complementary strands, and each strand is made up of small molecules called nucleotides. There are four types of nucleotides: adenine (A), cytosine (C), guanine (G), and thymine (T). Adenine and guanine are called purines, while pyrimidines consist of the other two types of nucleotides, cytosine and thymine. Adenine pairs with thymine and guanine pairs with cytosine by hydrogen bonds to form the double helical form. RNA is another nucleic acid in the cell, with also a four-letter alphabet of ribonucleotides {A, C, G, U}, where thymine (T) of DNA is replaced by uracil (U). Protein is another type of polymer that is made up of 20 amino acids. As put forward by Francis Crick [23] in 1958, the “central dogma” summarizes the information flow among the above biopolymers.

“The central dogma states that once ‘information’ has passed into protein it cannot get out again. The transfer of information from nucleic acid to nucleic acid, or from nucleic acid to protein, may be possible, but transfer from protein to protein, or from protein to nucleic acid, is impossible. Information means here the precise determination of sequence, either of bases in the nucleic acid or of amino acid residues in the protein.”

Since the advent of rapid DNA sequencing technologies in the late 1970s, the amount of data about the protein and DNA sequence of humans and other organisms has been growing at an exponential rate. By examining protein and DNA sequences, one can study the evolutionary relationships among species, populations, or genes; as well as infer the function of a new sequence based on the function of some known sequence that is a close neighbor in the evolution. The reconstruction of the evolutionary history of taxa is based on the assumption that the observed taxa sequences are homologous, or come from a common ancestor. Therefore the evolution that gives
rise to the observed data of the sequences under examination can be characterized by a tree structure.

The internal force of evolution lies in the interaction between mutation and heredity. Mutations are errors in DNA replication (Li, [77]). They can be classified by the type of change caused by the mutational event which includes

**substitutions:** the replacement of one nucleotide by another;

**deletions:** the removal of one or more nucleotides from the DNA;

**insertions:** addition of one or more nucleotides to the sequence;

**recombination:** crossing-over and gene-conversion;

**inversions:** the rotation by 180° of a double-stranded DNA segment consisting of two or more base pairs.

In this thesis (except in Chapter 5), we will focus only on the nucleotide data sets for which substitutions are assumed to be the only mutations. Therefore, the nucleotides at the same position of sequences to be studied are from a common ancestor, and this is called *positional homology*. As a result, we can introduce the concept of *site*, and the nucleotides of various taxa at the same site (position) are homologous and all have been evolved from a common ancestor nucleotide.

Nucleotide sequences are usually displayed in a matrix form. As an example, Table 1.1 displays the first 57 nucleotides of the 1382-nucleotide-long L1 gene from six human papillomaviruses and a rhesus papillomavirus (RhpV). This data set has been analyzed in [15, 14, 90, 101], and we will revisit it in Section 4.4.
Table 1.1: DNA sequences for a portion of the L1 gene for seven Group A9 papillomaviruses. See Section 4.4 for a detailed description of the data.

Li [77] and Waterman [126] provide excellent introductions to the evolution of biopolymers.

1.2 Phylogenetic Tree

A tree is a cycle-free connected graph. If all nodes of a tree are of degree three or fewer, the tree is called a binary tree. (Degree is the number of nodes connected to a node.) If all nodes have degree of either three or one, with a possible exception of at most one node with degree of two, the binary tree is called full binary tree or bifurcating tree. Those nodes with degree of one are external nodes or terminal nodes. The nodes with degree of three are internal nodes. If there exists a node with degree two, it is called the root node, and the tree is called rooted bifurcating tree. To understand why the bifurcating tree gets the name, suppose that the direction of the edges from parent node to daughter node(s) can be specified, then each node in a bifurcating tree is either an external node or it gives rise to two daughter nodes. If a tree has a node with degree of more than three and the node is connected to more than three nodes, the tree is called multifurcating tree, and is said to have star phylogeny in the phylogenetic context.
All life forms on earth share a common origin, and the relationship among a set of organisms is called a *phylogeny*. The relationship is usually described in a tree structure, or *phylogenetic tree*. From a mathematical point of view, a phylogenetic tree is a bifurcating tree with labeled external nodes. In addition, a true biological phylogeny should have a root to represent the common ancestor of the group of organisms. External nodes that are labeled represent the taxonomic units under study, and are referred to as *operational taxonomic units (OTUs)* by biologists. The taxonomic units can be species, populations, individuals or genes. Typically, the external nodes are all contemporaries, which is why a rooted phylogenetic tree is usually represented with external nodes at the bottom (or to the right) and the root node at the top (or to the left). A phylogenetic tree has two essential components: topology of the tree and the branch lengths. *Topology* of the tree refers to the branching pattern that connects the labeled OTUs, and the branch lengths are the lengths of the individual edges or branches. Topology receives most of the attention as it qualitatively characterizes the closeness of a pair of organisms relative to other pairs.

A phylogenetic tree is rooted when the location of the common ancestor is identified, or unrooted if the root is not located. Figure 1.1 displays an unrooted tree in (a) with five external nodes, labeled nodes 1, 2, 3, 4 and 5, and a congruent rooted tree in (b) for which the root is located on the edge between node 5 and internal node 8 – ancestor of nodes 1, 2, 3, and 4. A rooted phylogenetic tree is often taken to satisfy the assumption of a *molecular clock* if the rate of evolution is assumed to be constant over all edges. This entails the consequence that the sum of branch lengths for the path connecting any external node to the root node is equal.
Figure 1.1: Phylogenetic trees with five OTUs. (a) An unrooted tree. (b) A rooted tree with the root placed between nodes 5 and 8.

Tree counting is a classical combinatorics problem. The number of distinct unrooted topologies of $N \geq 3$ labeled external nodes is known since Schröder [105] in 1870 to be:

$$(2N - 5)!! = (2N - 5) \times (2N - 7) \times \cdots \times 3 = \frac{(2(N - 1))!}{2^{N-1}(N - 1)!},$$

where $n!!$ represent double factorial of an integer. This can be verified by induction. Since there are $2N - 3$ branches where a root can be placed and thus $2N - 3$ unique rooted trees result, it follows that the number of rooted phylogenies for $N$ external nodes is equal to the number of unrooted phylogenies for $N + 1$ external nodes, $(2N - 3)!!$. The number of distinct topologies grows very rapidly as the number of external nodes increases. By Stirling’s formula, we have an asymptotic approximation
of the double factorial

\[ (2N - 3)!! \sim \left( \frac{2}{e} \right)^{N-1}(N - 1)^{N-1} \sqrt{2}. \]

For \( N = 10 \), there are about two million unrooted trees; and for \( N = 20 \), about \( 2.2 \times 10^{20} \) of them and \( 8.1 \times 10^{21} \) rooted trees; and for \( N = 30 \) there are \( 8.7 \times 10^{36} \) distinct unrooted trees. The typical number of OTUs for phylogenetic analysis is between 10 and 50 in practice, for which the cardinality of possible phylogenies is enormous.

### 1.3 Phylogenetic Tree Reconstruction

Phylogenetic tree reconstruction is the process to derive a “best” phylogenetic tree, according to certain criteria, from the observed nucleotide sequence data at a group of OTUs. It is often referred to as inferring the phylogeny by biologists, possibly for the reason that the unobserved part of the tree for the ancestors is to be inferred. In mathematical terms, the aim is to find the tree that optimizes objective function(s) based on the criteria of comparing the trees. The following terms for tree building, tree (re)construction, estimation, and tree inference, will be used interchangeably. At the inception, phylogenetic tree construction used morphological data at OTUs. Most data sets analyzed in modern time are all molecular data, and nucleotide data is the focus of this thesis.

Phylogenetic tree construction is at the intersection of multiple disciplines, such as graph theory, computer science, systematics and statistics, and various methods and algorithms have been proposed in the literature. In general, the tree building methods can be classified into three schools based on the optimality criteria: distance-based criteria, maximum parsimony (MP) criteria and maximum likelihood (ML) criteria.
Distance-based methods first convert the similarity of data at pairs of OTUs into pairwise evolutionary distances. The distance can be *ad hoc* based on empirical knowledge, or it can be more systematically formulated. An example of the latter is the maximum likelihood distance (Felsenstein [39]) between two sequences using a stochastic model described in Section 1.4. Once the evolutionary distances between pairs of OTUs are determined, the original data at OTUs is no longer used. The problem now becomes a cluster analysis with the aim to form a binary tree so that the clustering of the nodes is congruent to the evolutionary distance. The most popular clustering methods are the UPGMA method by Sokal and Michener [107] and the neighbor-joining (NJ) method by Saitou and Nei [100]. UPGMA is an abbreviation for unweighted paired-group method with arithmetic mean, and it builds the tree from the external nodes and works upward iteratively in the phylogeny. It sequentially joins the two nodes with smallest distance, and then treats them as a new node and recalculates the distance from the new node to other nodes. The process continues until all of the nodes are joined. The NJ method starts with a star phylogeny with only one internal node, and sequentially strips off a pair of taxa from other taxa under the restriction that the sum of all branch lengths is minimal. This process continues until a bifurcating tree results. The UPGMA method is inspired by the assumption that the true tree follows the molecular clock, or the sum of branch lengths from any external node to the root node is the same. If the molecular clock condition does not hold, this method can fail to infer the correct tree topology. The NJ method was proved to reconstruct the true topology correctly if the distance used satisfies additivity (Studier and Keppler [112]) or additivity approximately holds (Atteson [5]). Given a tree, its edge lengths are said to be additive if the distance between
any pair of external nodes is equal to the sum of the lengths of all edges on the path connecting them.

Distance methods for tree building have become less popular recently, due in part to the nature of the deterministic process that cannot investigate other trees. In addition, the assumptions made are intermediate between those for ML and MP methods, and receive objections from both schools. Nevertheless, it is commonly used as the initial tree for ML tree building because of the fast execution of the process.

The maximum parsimony method is probably the most used of all three types of tree inference methods by biologists. It is based upon the principle of maximum parsimony, which is to find the tree that minimizes the total number of evolutionary steps (substitutions of nucleotide character in the context of DNA sequences) required to explain the observed data. Instead of directly building the tree, it assigns a cost to each tree, and the aim is to find the tree that minimizes the cost. The characters at the internal nodes are also determined such that they give the minimum cost function value for a given tree topology. Fitch [43] was among the first to apply the principle to data of nominal (unordered) characters (e.g., nucleotide and protein sequences), and his original criteria are termed traditional parsimony, for which the cost function is simply 1 if two characters are different and 0 if they are the same. Finding MP tree under the traditional parsimony criterion is equivalent to finding a minimal Steiner tree with the constraint of the tree being binary, under Hamming distance (the distance corresponding to the above defined cost function). The minimal Steiner tree problem has been well studied by computer scientists, and is NP-hard, i.e., no algorithm of polynomial time in the number of external nodes exist. Sankoff and Cedergren [103] generalized the cost function such that it depends on the values
of two characters, and the criterion is termed *weighted parsimony*. In their paper, they also provided an algorithm based on dynamic programming to compute the value of the cost function for a specific tree given the observed data at OTUs. The computation is more demanding for finding the MP tree under weighted parsimony criterion. It is worth noting that trees with the same topology but different branch lengths have the same cost function value under MP approach, and therefore, the MP method only reconstructs the tree topology and ignores the branch lengths.

The maximum likelihood method to reconstruct a tree is perhaps the easiest to understand for statisticians. It explicitly specifies a stochastic model for the change of characters over time, and most often the model is characterized by a continuous-time discrete-state time-reversible Markov process. Some stochastic models for nucleotide evolution are described in the next section. Given a tree topology and set of branch lengths, the likelihood of the observed data can be calculated by evaluating the probability of internal nodes taking on a particular combination of character values, and then summing over all possible combinations. In the 1950s, A. W. F. Edwards and L. Cavalli-Sforza first applied maximum likelihood approach to phylogeny construction using data of continuous variables ([28], [29] and [13]). Felsenstein ([34], 1981) introduced the basic algorithm for computing the likelihood of trees from data of discrete characters, and laid foundations for the likelihood methods most commonly used in molecular phylogeny today. Details of the maximum likelihood method for tree reconstruction are provided in Section 1.5.

Both the MP and ML approaches map a tree to an objective function value, and seek the tree with the optimal objective function score. There are various algorithms
which search the tree space and attempt to find the optimal tree based on the pre-
specified criteria and objective function. Exhaustive search of the entire space by
enumerating all possible trees is permissible if the number of OTUs is very small,
generally no more than 10 or so. Branch-and-bound type of algorithms can greatly
reduce the computational time if the objective function is monotone as an additional
taxon is added. It guarantees that the optimal tree is found, but it does not give infor-
mation about near-optimal trees. Exhaustive search and branch-and-bound methods
cannot be used in general once the number of taxa is greater than 20 because of the
prohibitive computational cost.

Numerous heuristic search methods have also been proposed as remedies, such
as branch-swapping, star decomposition, and divide-and-conquer approaches, when
the size of the problem does not allow finding the exact solution. Branch-swapping
methods are included in almost all popular phylogenetic analysis packages, e.g.,
PAUP* [113], PHYLIP [38], and fastDNAML [89]. They first build a tree with three
taxa among the OTUs. The rest of the taxa are added one at a time to optimize
the pre-specified criteria. During each addition, branch swapping is attempted to
improve on the objective function score. Branch swapping is to execute from a set of
pre-defined rearrangements to the tree topology, e.g., the local rearrangement transi-
tion strategy is a particular branch swapping strategy given in Kuhner, Yamato and
Felsenstein [69] for rooted trees. (See Figure 1.2 for illustration.) Analogous meth-
ods for unrooted trees are Nearest-neighbor interchanges (NNI), subtree pruning and
regrafting (SPR), and tree bisection and reconnection (TBR). After a tree is built
including all the taxa, thorough branch swapping is performed to determine whether
nearby trees can yield a better objective function score. The nearby trees are those
that can be reached by a few operations of branch swapping. The process ends when no nearby trees result in a better score. The star decomposition method is not limited to the use for neighbor-joining method with a distance matrix, but can be applied to any criterion for which the multifurcating trees can be evaluated. It starts with a star phylogeny with only one internal node, and the number of branches connected to this central node is reduced by one in each iterative step by pairing two of the nodes that result in the optimal objective function score. Yang’s PAML [134] program implemented this algorithm for tree searching. Algorithms based on the technique of divide-and-conquer have also been proposed in the literature of phylogenetic tree construction. It works by dividing the problem of inferring the optimal tree from *N* taxa into several problems of inferring optimal subtrees with smaller number of taxa. However, combining subtrees into the global solution could fail when the subtrees are not congruent. In sum, these heuristic search methods are all uphill search methods, with possible variants for slight improvement. An apparent drawback is that they tend to easily get stuck in the local optima and do not find the true global optimum solution. Unfortunately, the surface of the objective function is extremely complex and has numerous local peaks and valleys, even for a tree with moderate size. Though these methods may provide some time savings over the exact methods, the drawback of being trapped in the local optima is a serious problem.

Other methods that are intended to address the issues with uphill searching have also been proposed. Among them are the stochastic search methods of simulated annealing and genetic algorithms. Genetic algorithms (GA) have been developed for phylogenetic tree reconstruction, most notably by Lewis [75] for DNA sequences. The general genetic algorithm is inspired by the analogy of genetic transmission of
Figure 1.2: The local rearrangement transition strategy \[69, 76\]. A neighborhood is selected for local rearrangement (diagram (a)). The local rearrangement results in trees (b) or (c), in which the topology has been changed, or in tree (d), in which the topology stays the same.

Desirable characteristics to successive generations to the search for solutions that have certain optimal characteristics. GA’s represent the possible solutions as vectors (of attributes). Two vectors can possibly cross over, which result in the exchange of attributes. GA’s start with an initial set of vectors, and generate a new set of progeny of vectors according to a set of possible operations and rules about the survival of vectors. The objective function value is used to determine the fitness of a vector, or the probability of survival.

Simulated annealing, first postulated in the field of statistical mechanics in the 1950s \[83\], is a general optimization method for a discrete parameter space. It was adapted to the setting of phylogenetic tree reconstruction by a few authors \[6, 25, 81, 102\]. The algorithm of stochastic tree searching procedure is as follows: for the tree at Step \(i\), a proposal tree is derived by a certain branch-swapping strategy
described above. If the proposal tree results in a better objective function score, it is accepted as the tree at Step \( i + 1 \). If the proposal tree has a worse score, it can still be accepted with a probability depending on the difference of the scores and a pre-set sequence of parameters generally depending on \( i \), which is commonly called the cooling procedure. With the remaining probability, the tree at Step \( i + 1 \) remains the same as the tree at Step \( i \). Geman and Geman [46] showed that with the cooling schedule approaching zero but not too fast, the procedure will find a global optimum with probability one. In practice, the determination of an appropriate cooling schedule and the specification of the parameters within the cooling schedule are difficult and will cause the performance of the algorithm to vary greatly. The Stochastic Search Algorithm (SSA) method developed by Salter and Pearl [102] is one simulated annealing method that has reported success in finding the optimal trees with MP and ML criteria. In particular, they reported that their algorithm outperforms commonly used heuristic search methods as implemented in PAUP* and PHYLIP both in terms of time and its ability to locate the global optimal solution for the maximum likelihood criteria.

Another general class of methods that have recently been applied more and more often under the likelihood paradigm is the Bayesian phylogeny reconstruction [76, 82, 137, 63]. With a stochastic model of nucleotide evolution assumed, a prior distribution, most often a non-informative prior in a certain sense, is placed on the space of tree topology, branch lengths and other parameters in the evolution model. If the posterior distribution of tree conditional on data observed on the OTUs is found, the topology with maximum \( a \) posteriori probability (MAP, [137]) is considered the optimal tree. Since the posterior distribution is not tractable, Markov chain
Monte Carlo (MCMC) can be applied via the standard Metropolis-Hastings algorithm to draw dependent samples from the posterior distribution. In particular, the MrBayes package by Huelsenbeck and Ronquist [63] additionally implemented the method of Metropolis-coupled Markov chain Monte Carlo (MCMCMC) [47]. It runs \( n \) chains, and the \( i \)-th chain has the limiting distribution of \( f(x)\beta_i \), where \( f(x) \) is the desired posterior distribution of tree given data. \( \beta_i \) takes on the form of \( \frac{1}{1 + (i - 1)T} \) and \( T \) is a user-defined parameter. It follows that the first chain is a chain with the usual MCMC, and the other chains (heated chains) have limiting distributions with closer values of maximum and minimum. After all \( n \) chains add one step, a swap of states between two randomly chosen chains is attempted, however, the final sampling of the posterior distribution is based upon only the first chain. Since it is easier for a heated chain to visit the valleys of the likelihood surface, this algorithm has been reported to have more success at avoiding getting stuck in a local maximum for a long time and thus improve the mixing and performance of the chain. MCMC methods have the advantage that in theory the full posterior distribution of the trees and other parameters given data can be constructed and thus a credible set in Bayesian analysis can be formed as a measure of the high probability region. Compared to the traditional methods of ML tree reconstruction, it is computationally efficient to address certain questions of interest. However, as with any MCMC method for exploring a large and complex state space, convergence and mixing are of practical concern. Moreover, when the number of taxa is large, MCMC methods tend to give poor estimates of posterior probability of individual phylogenetic tree due to the enormous state space, and they may be computationally intensive. Huelsenbeck
et al. [62] presented a survey and discussed the major strengths and issues of the Bayesian phylogenetic inference methods via MCMC.

The three schools of tree building, distance-based, maximum parsimony and maximum likelihood, all make certain assumptions. Distance method in theory does not have to assume the independence between the sites, but almost all of the distance measures are summation of the distance over the individual sites. MP methods implicitly assume the independence of the columns in the data; while all practical ML methods make the assumption of independent sites to make the calculation of likelihood feasible. Distance-based methods assume that the pairwise distance measure derived from the data satisfies certain assumptions for the corresponding clustering algorithms. Maximum parsimony disallows homoplasy, or the reversal or back-substitution. Maximum likelihood methods explicitly specify the underlying stochastic evolution model for nucleotides, and relies on a full parametric approach. If the respective assumption holds, tree reconstructed based on each of the three criteria is statistically consistent, that is, data from infinite sites would give rise to the true tree. Distance methods do not use the original data once pairwise distances are obtained. MP methods do not use the data from all sites, e.g., sites with the same nucleotide do not discriminate among different trees and are discarded. In contrast, ML methods make use of all of the data. The assumptions for distance methods and MP methods, one can argue, are not realistic. If homoplasy does occur, the MP method will fail to find the true tree even with infinite number of sites. There are also ties among the three types of criteria: when the maximum likelihood distance derived from a stochastic evolutionary model is used to find pairwise distances, the distance methods can be viewed as a crude approximation to the ML criteria. Nonetheless, the drawback of the inability
to evaluate the stochastic model makes distance methods less desirable. While the ML method evaluates all possible combinations of ancestor states at internal nodes and uses the sum as the optimality criterion, the MP method considers only one set of ancestor states. When there is no common mechanism between sites, an MP approach can be regarded as an ML method [111]. In terms of computational cost, distance method costs the least time, and ML methods are the most expensive.

Earlier review by Swofford et al. [114] and chapters in books by Waterman [126], Lange [72], Durbin et al. [26] and Felsenstein [40] provide excellent surveys of the subject of phylogenetic trees and the three schools of tree buildings. Holmes [59] gave an account from a statistician’s perspective.

1.4 Stochastic Models of Nucleotide Evolution

In order to calculate the likelihood of a phylogenetic tree under the given data, the model that describes the probability of evolution from state (nucleotide) $i$ to state $j$ at a single site along a branch needs to be specified. The model is usually referred to as substitution model. A substitution model typically corresponds to a continuous-time four-state time-reversible Markov process [66]. The Markov property renders that the evolution at a site past a time point $t_0$ is independent of the history, given the state at $t_0$. The state space is the alphabet $\mathcal{A} = \{A, C, G, T\}$. The Markov model is characterized by a $4 \times 4$ infinitesimal transition matrix $Q$ with a number of parameters. Hence, the probability transition matrix over a fixed time length $t$ [66] is given by

$$P = e^{Qt}. \quad (1.1)$$
The above probability transition matrix determines the stationary distribution of each
nucleotide, $\pi_A$, $\pi_C$, $\pi_G$, and $\pi_T$, and in this case they are the limiting distribution
of the chain. Note that the term “transition” refers to the change of state in the
Markov process context, but it could also mean one of the two types of the nucleotide
substitution.

A Markov chain is time-reversible if at the stationary distribution, the probability
of going from state $i$ to $j$ in time $t$ is the same as going from state $j$ to $i$ in time $t$,
i.e.,

$$\pi_i P_{ij}(t) = \pi_j P_{ji}(t), \text{ for all } i, j, t. \quad (1.2)$$

Some commonly used time-reversible substitution models are described below.

**Jukes-Cantor Model**

The Jukes-Cantor (JC) model [65] is the simplest substitution model. It has only
one parameter, which describes the instantaneous rate of all nucleotide substitutions.
The infinitesimal transition matrix is given below:

$$
\begin{pmatrix}
\frac{3}{4}\mu & \frac{1}{4}\mu & \frac{1}{4}\mu & \frac{1}{4}\mu \\
\frac{1}{4}\mu & \frac{3}{4}\mu & \frac{1}{4}\mu & \frac{1}{4}\mu \\
\frac{1}{4}\mu & \frac{1}{4}\mu & \frac{3}{4}\mu & \frac{1}{4}\mu \\
\frac{1}{4}\mu & \frac{1}{4}\mu & \frac{1}{4}\mu & \frac{3}{4}\mu
\end{pmatrix},
$$

where the columns and rows represent the four states in the alphabet $\mathcal{A} = \{A, C, G, T\}$
in that order. This convention applies also to the infinitesimal transition matrices in
other models that follow in this section. Therefore, the probability from nucleotide $i$
to \( j \) in time \( t \) for JC model is:

\[
P_{ij}(t) = \frac{1}{4} - \frac{1}{4} e^{-\mu t} + \delta_{ij} e^{-\mu t}, \tag{1.3}
\]

where \( \delta_{ij} \) is the Kronecker delta function that has value 1 when \( i = j \) and 0 otherwise. Note that the stationary probability of each nucleotide as determined by the model is 1/4.

**Kimura’s Two-Parameter Model**

Kimura [67] proposed a two-parameter model (K2P). There are two types of nucleotide substitutions: transitions (changes of A ↔ G and C ↔ T) and transversions (changes of A ↔ C, A ↔ T, G ↔ C, and G ↔ T). One parameter \( \mu \) governs the overall rate of substitution, and another parameter \( \kappa \) models the transition/transversion ratio. The infinitesimal transition matrix is given by

\[
\begin{pmatrix}
-\frac{1}{4}\mu(\kappa + 2) & \frac{1}{4}\mu & \frac{1}{4}\mu\kappa & \frac{1}{4}\mu \\
\frac{1}{4}\mu & -\frac{1}{4}\mu(\kappa + 2) & \frac{1}{4}\mu & \frac{1}{4}\mu\kappa \\
\frac{1}{4}\mu\kappa & \frac{1}{4}\mu & -\frac{1}{4}\mu(\kappa + 2) & \frac{1}{4}\mu \\
\frac{1}{4}\mu & \frac{1}{4}\mu\kappa & \frac{1}{4}\mu & -\frac{1}{4}\mu(\kappa + 2)
\end{pmatrix}.
\]

The transition probability over time \( t \) is

\[
P_{ij} = \frac{1}{4} - \frac{1}{4} e^{-\mu t} + \frac{1}{2} \left( e^{-\mu t} - e^{-\mu t(\frac{\kappa + 1}{4})} \right) \epsilon_{ij} + e^{-\mu t(\frac{\kappa + 1}{4})} \delta_{ij}, \tag{1.4}
\]

where \( \delta_{ij} \) is the Kronecker delta function, and \( \epsilon_{ij} \) is 0 if the change from \( i \) to \( j \) is a transversion and 1 otherwise. Note that the stationary distribution of each nucleotide is determined to be 1/4, and that when \( \kappa = 1 \) the Kimura model reduces to the Jukes-Cantor model.
Hasegawa, Kishino and Yano’s Five-Parameter Model

Hasegawa, Kishino and Yano [53] added another three parameters to model the stationary distribution of the four nucleotides to extend the K2P model. The infinitesimal transition matrix of this five-parameter model (HKY85) is given by

\[
\begin{pmatrix}
-\mu(k\pi_G + \pi_Y) & \mu\pi_C & \mu k\pi_G & \mu\pi_T \\
\mu\pi_A & -\mu(k\pi_T + \pi_R) & \mu k\pi_A & \mu\pi_T \\
\mu k\pi_A & \mu k\pi_C & -\mu(k\pi_A + \pi_Y) & \mu k\pi_T \\
\mu k\pi_A & \mu k\pi_C & \mu k\pi_T & -\mu(k\pi_C + \pi_R)
\end{pmatrix},
\]

where \( \pi_j \) is the equilibrium frequency of nucleotide \( j \), \( \pi_R = \pi_A + \pi_G \) and \( \pi_Y = \pi_C + \pi_T \).

The probability of nucleotide \( i \) evolving into \( j \) over time \( t \) is given by

\[
P_{ij}(t) = \pi_j - \pi_j e^{-\mu t} + \frac{\pi_j}{\Pi_j} (e^{-\mu t} - e^{-\mu t(1+\Pi_j(k-1))})\delta_{ij} + e^{-\mu t(1+\Pi_j(k-1))} \delta_{ij},
\]

where \( \Pi_j = \pi_R = \pi_A + \pi_G \) if \( j \) is a purine and \( \Pi_j = \pi_Y = \pi_C + \pi_T \) if \( j \) is a pyrimidine. Note that by setting \( \pi_i = \frac{1}{4} \) for all \( i \), HKY85 model reduces to the K2P model, and additionally letting \( k = 1 \) gives the Jukes-Cantor model.

Felsenstein’s Five-Parameter Model

Felsenstein [36] also proposed a five-parameter model (F84 model) as a generalization of Kimura’s model to allow for unequal equilibrium frequencies of the four nucleotides. The F84 model and HKY85 model are not equivalent models, because the space that the infinitesimal probability matrix spans for each model is different. In the special case where the equilibrium frequencies for purines and pyrimidines are equal, the two models are equivalent. The idea behind the F84 model is that the substitution process can be thought of as having two separate components: a general substitution rate which produces both transitions and transversions, and a
within-group substitution rate which produces only transitions. The instantaneous rate matrix for this model is shown below,

\[
\begin{pmatrix}
- & \mu \pi C & \mu \pi G \left(1 + \frac{K}{\pi R}\right) & \mu \pi T \\
\mu \pi A & - & \mu \pi G & \mu \pi T \left(1 + \frac{K}{\pi Y}\right) \\
\mu \pi A \left(1 + \frac{K}{\pi R}\right) & \mu \pi C & - & \mu \pi T \\
\mu \pi A & \mu \pi C \left(1 + \frac{K}{\pi Y}\right) & \mu \pi G & -
\end{pmatrix},
\]

where \( \pi_j \) is the equilibrium frequency of nucleotide \( j \), \( \pi_R = \pi_A + \pi_G \), \( \pi_Y = \pi_C + \pi_T \), and the diagonal elements are the negatives of the sums of the corresponding rows.

From the instantaneous rate matrix, we can calculate the transition probability as follows:

\[
P_{ij}(t) = \pi_j - \pi_j e^{-\mu t} + \frac{\pi_j}{\Pi_j} (e^{-\mu t} - e^{-\mu t(K+1)}) \epsilon_{ij} + e^{-\mu t(K+1)} \delta_{ij},
\]

where \( \Pi_j = \pi_R \) if \( j \) is a purine and \( \Pi_j = \pi_Y \) if \( j \) is a pyrimidine. We note that setting \( \pi_i = \frac{1}{4} \) for all \( i \) gives the K2P model and additionally letting \( K = 0 \) gives the JC model.

More General Models

The most general substitution model for characterizing a time-reversible Markov process of the evolution of a nucleotide is the general time-reversible (GTR) model that requires 10 parameters, and the infinitesimal transition matrix is as follows ([71], [96]):
\[
\begin{pmatrix}
- & \mu \pi C^\alpha AC & \mu \pi C^\alpha AG & \mu \pi T^\alpha AT \\
\mu \pi A^\alpha AC & - & \mu \pi C^\alpha CG & \mu \pi T^\alpha CT \\
\mu \pi A^\alpha AG & \mu \pi C^\alpha CG & - & \mu \pi T^\alpha GT \\
\mu \pi A^\alpha AT & \mu \pi C^\alpha CT & \mu \pi C^\alpha GT & - \\
\end{pmatrix},
\]

where the diagonal elements are the negatives of the sums of the corresponding rows.

All of the previously described substitution models are special cases of the GTR model. However, when the number of parameters increases, the issue of over-parameterization arises. It seems the five-parameter models of F84 and HKY85 are sufficient for many of the data sets in practice.

All of the aforementioned models described thus far are those for a single site. It has long been recognized that substitutions occur much more often at some sites than others. A common remedy is to place a prior on a site-specific overall substitution rate parameter, denoted by \( r \). Yang proposed using a gamma distribution [128] and the discretized version of gamma distribution [129], for which the mean in the family of gamma distributions is fixed at 1 and thus only a single parameter is to be estimated. Felsenstein and Churchill [41] proposed an algorithm similar to Yang’s, but in a Hidden Markov Model format. In general, suppose that the prior has c.d.f. of \( \Psi \) that may depend on some parameter(s) (for illustrative purpose, we assume a proper prior here), and also that the substitution model at a site with \( r = 1 \) has an instantaneous transition matrix of \( Q \). Then the probability transition matrix over time \( t \) is

\[
P(t) = E_r e^{rQ^t} = \int e^{rQ^t} \, d\Psi(r).
\]

Evolutionary models with gaps to model in/del (insertions and deletions) have also been proposed. A simplest model would be to extend the state space by including a state of ‘-’ to represent a gap for a single site in/del. This is critiqued for making
the in/del occur in length one and independently, and not allowing the gaps to occur in blocks. Numerous attempts have been made to improve on this aspect. Allison et al. [3] introduced an affine-type penalty. Thorne et al. [118] proposed a model of fragment substitution, but it has only been applied to the case of two sequences. Mitchison & Durbin [84] modified the Hidden Markov Model (HMM) of Krogh et al. [68] and applied it to this context to allow an affine-type gap penalty, but they reported that it could handle only a small number of sequences. Regarding the above attempts, once the independence of the columns in the data is not assumed in order to model in/dels of length more than one character, the computation of the likelihood for the data from the sequences is no longer a simple product of likelihood at each site, and thus greatly increases the computational burden.

1.5 Maximum Likelihood Estimation

We will first illustrate how the likelihood of a rooted tree is calculated given the nucleotide data of the external nodes at a single site once the substitution model is specified. This was described in the pioneering papers by Felsenstein [34, 35].

Consider a rooted tree with $N$ external nodes, $b = (b_1, b_2, \ldots, b_N)$. Denote the root node as $v_0$, and the other internal nodes $v_1, \ldots, v_{N-2}$. Let $\delta(v_i)$ denote the set of the two daughter nodes of an internal node (possibly the root node) $v_i, i = 0, \ldots, N - 2$. Suppose that $w_i$ is a daughter node of $v_i$, and in addition let $t(v_i, w_i)$ denote the length of the edge connecting $v_i$ and $w_i$. At a generic site, the nucleotide at a node $v_i, i = 0, \ldots, N - 2$, is denoted by $a(v_i)$, and furthermore, $(a(v_0), \ldots, a(v_{N-2}))$ is denoted by the vector $a$. Let $a(b_i)$ denote the nucleotide at an external node $b_i, i = 1, \ldots, N$. By summing over all possible nucleotide configurations at the internal
nodes, we determine the likelihood of observing the data \( d = (a(b_1), \ldots, a(b_N)) \) at the external nodes under the given tree as

\[
L(d) = \sum_{a} \pi(a(v_i)) \prod_{i=0, \ldots, N-2} \prod_{w \in \delta(v_i)} P_{a(v_i), a(w)}(t(v_i, w)),
\]

(1.7)

where \( \pi(a) \) is the equilibrium frequency of nucleotide \( a \).

As an example, Figure 1.3 depicts a rooted tree with four external nodes, which also satisfies the molecular clock assumption, i.e., the distance from the root node to any of the four external nodes is equal. The likelihood at a site, following Equation (1.7), is

\[
L(d) = \sum_{(c,d,e) \in A^3} \pi_c \{ P_{c,d}(t_1) (P_{d,a(b_1)}(t_2) P_{d,a(b_2)}(t_2)) \} \{ P_{c,e}(t_3) (P_{e,a(b_3)}(t_4) P_{e,a(b_4)}(t_4)) \}.
\]

Figure 1.3: A rooted tree with four external nodes \( b = (b_1, b_2, b_3, b_4) \), which satisfies molecular clock assumption.

The sum in Equation (1.7) involves approximately \( |A|^{N-1} \) terms, and the implementation is not computationally feasible as \( N \) gets moderately large. Felsenstein [34] described an algorithm referred to as the peeling or pruning algorithm, which greatly reduces the number of terms in the summation.

To illustrate the idea, consider an internal node \( v_j \) and a generic site, and the peeling algorithm makes use of the conditional independence of the parent node and
the daughter nodes of \( v_j \) given the nucleotide at \( v_j \), which follows from the Markov property. The likelihood of a subtree can be aggregated into its ancestor node at the top, and by pruning the tree from bottom up, we calculate the likelihood by a traversal of the internal nodes. Denote by \( f_{u_1}(a) \) the probability of having a nucleotide \( a \in \mathcal{A} \) at an internal node \( u_1 \). Suppose that \( u_2 \) and \( u_3 \) are the two daughter nodes, and suppose that \( f_{u_2}(a) \) and \( f_{u_3}(a) \) are known for all \( a \in \mathcal{A} \). Then, the following holds true, and it reveals the recursive relationship when traversing the nodes from bottom up along the phylogeny.

\[
f_{u_1}(a) = \sum_{(c,d) \in \mathcal{A}^2} f_{u_2}(c) P_{ac}(t_{u_1 u_2}) f_{u_3}(d) P_{ad}(t_{u_1 u_3}).
\]

(1.8)

At the external node \( b_i \) which has an observed nucleotide \( a_i \) at the site of interest, the initial condition is given by

\[
f_{b_i}(a) = \begin{cases} 1, & \text{if } a = a_i; \\ 0, & \text{otherwise.} \end{cases}
\]

The final likelihood is \( \sum_{a \in \mathcal{A}} \pi(a) f_{v_0}(a) \), where \( v_0 \) is the root node.

With the nucleotide data of the sequences at external nodes from \( L \) independent sites, the log-likelihood is simply the sum of the single-site log-likelihood over all the sites. Whereas the tree shown in Figure 1.3 satisfies the molecular clock assumption, the pruning algorithm works for any bifurcating tree with or without this assumption.

The time-reversibility property of the substitution model for nucleotide evolution allows the root node of the tree to be placed anywhere in the tree and the resulting likelihood to remain equal, which was termed the pulley principle by Felsenstein [34]. Therefore, the likelihood of an unrooted tree is thus well-defined and is calculated
as that of the resulting rooted tree with the root placed arbitrarily along the tree branches.

So far we have introduced how the likelihood can be evaluated for a given tree topology with given branch lengths under a specified substitution model on nucleotide evolution with a given set of evolutionary parameter values. In the remainder of this thesis, we use the word “parameter” to refer to both the evolutionary parameter in the substitution model and the collection of the three unknown components: tree topology, branch lengths and evolutionary parameters, if no confusion is likely based on the context. In theory, if we can find the value for the set of tree topology, branch lengths and evolutionary parameters that maximizes the likelihood, it is the maximum likelihood estimator (MLE). The tree topologies form a finite discrete space with very large cardinality, while branch lengths can be thought of as a vector with constraints congruent with the specified topology and bifurcating tree structure. The algorithms for ML criteria in Section 1.3 discussed only the search through the discrete space of tree topologies. Once an optimal topology is found, ideally a multi-dimensional optimization should be performed on the branch lengths vector with proper constraints to find the joint MLE, but the computational burden makes it not feasible to implement. Almost all packages employ the following procedure: apply the Newton-Raphson type of algorithm or its derivative-free variant to optimize the branch lengths one at a time, and repeat the process until the whole vector converges. This uphill optimization method by optimizing one branch length at a time is susceptible to being stuck at a local maximum.

The discussion of the estimation of evolutionary parameters is deferred to the next section.
1.6 Evolutionary Parameters Estimation

Ideally, in order to infer the ML tree, we would search simultaneously for globally optimal values of tree topology, branch lengths vector and the parameters in the evolutionary model over the entire parameter space, that is, to consider every possible tree topology and optimize jointly all numeric parameters under each tree, and choose the resulting tree(s) with the highest likelihood. For a given tree, a multi-dimensional optimization using Newton-Raphson method could be performed in theory. However, the high computational cost of such method renders it not feasible in practice. Yang [131] cited previous work and reported a simulation study on a four-taxon example and suggested that parameter estimates would be fairly stable across tree topologies as long as the topology is not “too wrong”. That leads to the common implementation that estimates the model parameters on some reasonably good tree and then fixes the resulting estimates in a search for better trees under the specified model. Successive iteration between estimation of numeric parameters and searching within the discrete space of tree topologies is carried out until the process converges.

In this section, we will review how the parameters in the substitution models are estimated by popularly used packages. We only discuss substitution models with five parameters (F84 or HKY85) under the assumption of homogeneous rate across sites. The four nucleotide frequency parameters, \( \pi_i, i = A, C, G, \text{ and } T \), are typically computed at the beginning of the run as the empirical frequencies from the observed data, e.g.,

\[
\pi_A = \frac{\text{number of nucleotide } A \text{ in the } N \text{ sequences at external nodes}}{N L},
\]
where \( N \) is the number of sequences or the external nodes and \( L \) is the number of sites in each sequence.

Heuristically, these empirical frequencies should not be far from the MLE, and many packages including PHYLIP and PAUP* use them as the final estimates. In SSA, an option is allowed to recalculate the ML estimators under a tree with the topology, branch lengths vector and other evolutionary parameters fixed. In practice, one is generally content with values of the nucleotide frequency parameters fixed at the values of the empirical estimates during the search for MLE of other parameters and tree. It is a practical compromise – the model containing only two parameters is relatively easy for implementation of optimization without incurring too much of the computational burden.

Now we will focus on the remaining two parameters, \( \mu \) and \( K \). First, it is important to note that the parameter \( \mu \) and the entire \( (2N - 2) \)-dimensional branch lengths vector for a rooted tree with \( N \) OTUs are not identifiable. This is easily seen from the form of the infinitesimal transition probabilities which always has \( \mu \) and \( t \) together as a product, and therefore, a constant \( c \) times \( \mu \) and all branch lengths divided by the same \( c \) yield the same probability matrix. To make the parameterization identifiable, one parameter must be dropped, and we assume throughout this thesis that one of the branch lengths is fixed and thus \( \mu \) is to be estimated.

There are different parameterizations of the other parameter, which appears as \( K \) in F84 model, \( \kappa \) in HKY85 model, and transition/transversion ratio \( R \) in certain other software implementations. We first review the traditional \textit{ad hoc} estimate of transition/transversion ratio not within the framework of maximum likelihood simultaneous estimation of the parameter with the tree. For any given two homologous
sequences, it is easy to count the apparent transitions and transversions based on the different nucleotides and find their ratio. Clearly these apparent numbers of mutations undercount the true numbers due to the possible multiple mutations along the path of evolution, but they provide a first-order approximation and taking the ratio may diminish the effect of this underestimation for each of the numerator and denominator. Following this approach, the ratio estimates based on pairs of sequences can be found for more than two sequences, but it is not clear how to combine these estimates into one under a phylogeny. Alternatively, if parsimony is used and the nucleotides at the ancestor (internal) nodes are found, the number of “apparent” transitions and transversion (once again, they are underestimates) can be used to derive the overall ratio. Many papers have explored methods along the lines of the above two approaches, and some made attempts to give ad hoc improvement on the shortcoming for underestimation of the number of mutations. Wakeley [122] provided a survey of the above two types of approaches, and in addition he [122, 123] pointed out the underestimation of the transition/transversion ratio was most severe for sequences that diverged from the ancestor early and that heterogeneity among sites could also be an important source for underestimation. Pollock and Goldstein [91] proposed a weighted average of the estimates based on all pairs of sequences. Assuming the phylogeny is known, Purvis and Bromham [92] proposed a scheme of iteratively reweighted least-squares regression to estimate the instantaneous transversion rate for a data set. Ina [64] applied Haldane’s correction to the ratio estimate for a pair of sequences, and provided an estimate of the variance as well. Yang and Kumar [135] proposed a modification to correct the observed numbers of transitions and transversions under the parsimony
approach that builds the MP tree and assigns the most parsimonious residues at the ancestor nodes.

Despite the relatively higher computational cost, the ML method of estimating the transition/transversion ratio parameter has also been discussed in the literature and implemented in some software packages [134, 94, 113]. As will be seen in the remainder of the thesis, the MLE of evolutionary parameters enjoys good statistical properties of consistency and efficiency.

The SSA method also estimates the MLE of evolutionary parameters $\mu$ and $K$, along with tree topology and branch lengths vector. Because of the fact that the simulated annealing algorithm is not an uphill search, it allows the proposed states for parameters and trees during the search process to explore areas of the likelihood surface other than just an individual peak. Therefore, as is reported, SSA tends to be less likely to find only a local maximum.

PAML [134] is the only package that is known to also give a measurement of the precision of the estimate, or a “standard error”. Although never proving the efficiency results, Yang inverted the Hessian matrix while implementing the ML search and claimed that it would give the estimate of the variance. He termed this method of variance estimation as the curvature method, following perhaps R. A. Fisher's original terminology when referring to the observed Fisher information. In the next chapter, the asymptotic efficiency property will be formally established and an alternative variance estimate is proposed and compared to the observed Fisher information.

In the literature, many authors have commented that the maximum likelihood approach of estimating tree topology is somewhat robust to the specification of values of parameters, and even to the type of substitution model [52, 64, 131, 138, 136].
However, the correct ML estimation of evolutionary parameters is important to ensure that the statistical properties hold. In the example in Chapter 4, it will be demonstrated that a default value of transition/transversion ratio parameter leads to a non-optimal tree topology.
CHAPTER 2

ASYMPTOTIC EFFICIENCY OF EVOLUTIONARY PARAMETERS WHEN TREE IS KNOWN

In Chapters 2 through 4 of this thesis, the nucleotide sequences that will be studied are assumed to be already aligned. Therefore, the positional homology is given by column of the aligned sequences, referred to hereafter as a site. Moreover, we will assume that the sequences all consist of the four types of nucleotides and no gaps are allowed. In most of the current phylogenetic tree building schemes, gaps are not allowed or taken to be uninformative, since no models to handle in/del events have been commonly accepted.

We begin the Chapter by introducing the notations and showing the consistency property of maximum likelihood method in general for estimating evolutionary parameters and the phylogeny as a corollary of known results. Then we will show the new result of asymptotic efficiency of the maximum likelihood estimators of evolutionary parameters with tree fixed. Next a Monte Carlo estimator to the expected Fisher information is proposed to approximate the asymptotic variance-covariance matrix and estimate the variability associated with the MLE. In Section 2.4, simulation results are presented to compare the performance of this method to that of observed Fisher information, which has been implemented in a phylogenetic analysis
package. The new proposed Monte Carlo estimator is demonstrated to have superior performance.

In this Chapter, except in Section 2.1, the evolutionary relationship, including both the tree topology and branch lengths, among the sequences to be studied is assumed to be fixed and known, unless stated otherwise.

2.1 Background

2.1.1 Notation

Suppose that there are $N$ sequences, and each consists of $L$ nucleotides. Typically, the data of the sequences to be studied is presented as an $N \times L$ matrix. Each row corresponds to a sequence, and at each of $L$ sites (columns) the entries of the matrix take on values from the four-nucleotide alphabet $\mathcal{A} = \{A, C, G, T\}$. (See Table 1.1 for an illustration.) In the discussions that follow, we will consider a different representation of data. At each site, for a column of the matrix representation, there are $4^N$ unique patterns of the nucleotide configurations. The data, $\mathbf{X}$, is now alternatively displayed as the vector of cell counts for the $4^N$ site patterns, e.g., in Table 1.1 site pattern $(A, A, A, A, A, A)$ appears 7 times, and the respective cell count for this entry in the $4^7$-dimensional vector is 7.

We denote the tree topology by $T$, and the vector of branch lengths for the tree by $\mathbf{t}$. A substitution model (Section 1.4) is used to describe the underlying mechanism of nucleotide evolution at a single site. Three assumptions associated with the evolution across the sites are invoked:

**Assumption 1** All sites evolve according to the same tree.

**Assumption 2** All sites evolve independently.
Assumption 3 All sites evolve following the same substitution model with the same parameter value.

With these three assumptions, given the total number of sites $L$ for the sequences, the random $N$-tuples of nucleotide configurations at the external nodes of the tree at a site are independent and identically distributed, and thus the cell counts data $X$ has a multinomial distribution. The probability associated with the $i$-th cell or the $i$-th site pattern, $p_i$, is determined by the Markov process of nucleotide evolution and the underlying phylogenetic tree.

Note that the cell counts representation of data usually has many empty cells and is not an economic way to present the data. However, the multinomial distribution of $X$ allows classical statistical theory to be applied for studying the properties of the maximum likelihood estimator.

2.1.2 Consistency of ML Tree

Chang [17] cited the previous work by Chang and Hartigan [18] and Steel et al. [109, 110] that under mild conditions, the method of maximum likelihood has the statistical property of consistency for estimating tree topology. In addition, he [17] showed that the maximum likelihood estimator for the parameters that determine the edge transition probabilities, if the family of edge transition probability matrices satisfies certain conditions, is consistent. An edge refers to a branch that connects a pair of nodes. Chang did not specify the form of substitution model in the paper, but rather worked with the full models in terms of the transition probabilities of the Markov process at each edge.
Suppose that there are $N$ labeled external nodes and that they are connected with an underlying unrooted tree structure. At each of the external nodes, we observe residuals at the $L$ sites. Residuals are from a finite state space with cardinality $C$. Changes between states follow a continuous-time Markov process. At each edge of the tree, an edge probability transition matrix (at most $C^2$ parameters) characterizes this Markov process of residual state changes. Let $X_i$ represent the random $N$-tuple of residuals at the $i$-th site of the $N$ external nodes. Suppose that the observed data $X_1, X_2, \ldots, X_L$ are independent and identically distributed given the total number of sites $L$ and it follows a parametric distribution that is fully specified by the parameter $\theta$, which is the collection of tree topology, marginal probabilities for the states at a specified node, and the entries in the edge transition matrices. From the observed data, the parameter is estimated by ML approach. Chang proved the fact (Theorem 5.1 in [17]) that the maximum likelihood estimate of parameter, $\hat{\theta}_L$, converges in probability to the true parameter as the number of sites $L$ approaches infinity if the following three conditions hold.

1. Each tree in the tree space has a non-degenerate topology, that is, tree space is restricted to bifurcating trees only.

2. The marginal probabilities of all $C$ character states are positive for some node.

3. The edge transition matrices are invertible, not equal to a permutation matrix, and belong to a class of matrices that is strongly reconstructible from rows.

The class of matrices that are strongly reconstructible from rows is defined as follows:
Definition 2.1.1 A set of matrices $\mathcal{M}$ is strongly reconstructible from rows if, for each $M \in \mathcal{M}$ and each permutation matrix $R \neq I$, we have $RM \notin \overline{\mathcal{M}}$.

The closure $\overline{\mathcal{M}}$ is defined in the space of $C \times C$ matrices.

Based on Chang’s Theorem above, we will show that for the common nucleotide substitution models, introduced in Section 1.4, including the F84 and HKY85 models, the MLE of tree topology, the branch lengths vector and the evolutionary parameters is consistent. As an example, let’s consider Felsenstein’s F84 model. Now we will verify the three conditions of Chang’s Theorem and thus establish the consistency of the MLE. Under the F84 model, Equation (1.6) of transition probabilities between two nucleotides can be rewritten as

$$P_{ij}(t) = \begin{cases} \pi_j + \pi_j \left( \frac{1}{\Pi_j} - 1 \right) e^{-\mu t} + \left( \frac{\Pi_j - \pi_j}{\Pi_j} \right) e^{-\mu(t+1)}, & \text{if } i = j; \\ \pi_j + \pi_j \left( \frac{1}{\Pi_j} - 1 \right) e^{-\mu t} - \left( \frac{\pi_j}{\Pi_j} \right) e^{-\mu(t+1)}, & \text{if } i \neq j, \text{ transition}; \\ \pi_j (1 - e^{-\mu t}), & \text{if } i \neq j, \text{ transversion}, \end{cases}$$

(2.1)

where $\Pi_j = \pi_A + \pi_G$ if nucleotide $j$ is a purine (A or G) and $\Pi_j = \pi_C + \pi_T$ if nucleotide $j$ is a pyrimidine (C or T).

Condition 1 holds true because we only consider bifurcating trees. Note that the equilibrium nucleotide frequencies are assumed to be all positive in F84 model, and thus Condition 2 is satisfied. An edge transition matrix is invertible, because any edge transition matrix can be represented as $e^{Qt}$, where $Q$ is an infinitesimal probability matrix as shown in Equation (1.1), and $t$ is the branch length of the edge of interest. Thus,

$$\det(e^{Qt}) = e^{\text{Trace}(Q)t} > 0.$$  

(2.2)
The equality in (2.2) holds true because if we denote the eigenvalues of $Q$ as $\lambda_1, \ldots, \lambda_4$ (one of which must be 0), and let $U^{-1}$ be the matrix whose rows are formed by the normalized left eigenvectors, then

$$e^{Qt} = U \begin{pmatrix} \exp(\lambda_1 t) & 0 & 0 & 0 \\ 0 & \exp(\lambda_2 t) & 0 & 0 \\ 0 & 0 & \exp(\lambda_3 t) & 0 \\ 0 & 0 & 0 & \exp(\lambda_4 t) \end{pmatrix} U^{-1}.$$

For any finite edge length $t$, the edge probability matrix is not equal to a permutation matrix as none of the diagonal entries can be 0. Note that for $i, j = 1, \ldots, 4$, over any length $t$, the transition probabilities satisfy the property of

$$P(j, j) > P(i, j),$$

which can be easily verified from (2.1). In other words, the probability of observing the same nucleotide $j$ at both ends of an edge is greater than that of observing a change of nucleotide from $i$ to $j$. With the strict inequality, the set of edge transition probability matrices induced by the F84 model is strongly reconstructible from rows.

By Chang’s Theorem, the MLE of $\theta$ defined as the set encompassing the tree topology, nucleotide marginal probabilities at a node and entries of all of the edge transition probability matrices is consistent. It implies that the tree topology and the three nucleotide equilibrium frequency parameters in F84 model are consistent. Based on the form of transition probabilities in (2.1), it follows that, up to a scaling factor for parameter $\mu$, the edge probability matrices uniquely determine the branch length vector and the parameters $\mu$ and $K$. So long as the parameterization of $\mu$ and branch lengths vector satisfy the identifiability condition, they are consistent as well.

This leads to the following Corollary
Corollary 2.1.1 Under the F84 model, under the conditions that the tree is a bifurcating tree and the branch lengths vector and $\mu$ are identifiable, the maximum likelihood estimator of tree topology, branch lengths and evolutionary parameters is consistent.

The above reasoning applies to other nucleotide substitution models, such as Jukes-Cantor, K2P, and HKY85 models. In general, it would seem that any biologically plausible substitution model should satisfy the property that the diagonal value of the edge transition probability matrix is greater than any other value in the same column. Thus, under that family of models, the maximum likelihood estimator will be consistent if the edge probability matrix can uniquely determine the parameters in the model.

Independent of Chang’s work, Rogers [98] gave a proof of the identifiability of the tree topology for continuous-time time-reversible Markov models, which implies that the MLE of tree topology is consistent. Rogers [99] adapted Wald’s proof for consistency of the MLE for continuous parameters to the case of ML estimation for tree topology and showed that the topology estimator is consistent in the general class of GTR model for which substitution rates vary according to the invariable sites plus Gamma distribution.

2.2 Asymptotic Efficiency of MLE of Evolutionary Parameters

In the last section, it has been shown that the MLE of the tree topology, branch lengths, and the evolutionary parameters is consistent under the F84 model. However, consistency is a rather weak statistical property of an estimator. There are confusions
regarding whether the ML estimator of the continuous parameters other than the
discrete parameter of tree topology possesses the efficiency property [130, 132, 133].
In this section, assuming that the true tree topology and the branch lengths are
known, we will use the classical statistical theory on maximum likelihood estimation in
multinomial settings to prove the efficiency property for the MLE of the evolutionary
parameters. In Chapter 3, the general results of efficiency will be established when
topology, branch lengths and evolutionary parameters are simultaneously estimated
by the maximum likelihood method.

2.2.1 Birch’s Conditions

We begin with describing the settings of classical multinomial maximum like-
lihood estimation and Birch’s conditions. The requirement of regularity conditions for
efficiency results for the MLE in classical theory for multinomial sampling is different
from the usual conditions for \textit{i.i.d.} cases. The conditions first given by Birch [9] are
more general conditions, but the set of conditions [11] stated below is commonly used
and is usually sufficiently general for practical purposes.

Let us first fix notations. Let \( \mathbf{X} = (X_1, \ldots, X_T)' \) denote a multinomial random
vector of cell counts with the vector of multinomial cell probabilities \( \mathbf{p} = (p_1, \ldots, p_T)' \)
satisfying \( \sum_{i=1}^{T} p_i = 1 \), and sample size \( L = \sum_{i=1}^{T} X_i \). Note that \( T \) is equal to \( 4^N \)
for the dimension of the multinomial cell probability vector for a tree with \( N \) taxa.
Assume that \( p_i > 0 \) for each \( i \). The distribution of \( \mathbf{X} \) is given by
\[
\Pr \{ \mathbf{X} = (x_1, \ldots, x_T)' \} = L! \prod_{i=1}^{T} \frac{p_i^{x_i}}{x_i!} \tag{2.3}
\]
for non-negative integers \( x_i, 1 \leq i \leq T \), satisfying \( \sum x_i = L \). Parametric models for
\( \mathbf{p} \) are specified as \( \mathbf{p} = \mathbf{f}(\theta) \) for an \( s \)-dimensional \( s < T - 1 \) vector \( \theta \) in an open
set $\Theta \subset \mathbb{R}^S$. An arbitrary “true” $p$ is denoted by $\pi = f(\varphi)$, where $\varphi$ is the true parameter value. Birch’s conditions are listed as follows.

1. The point $\varphi$ is an interior point of $\Theta$, so that $\varphi$ is not on the boundary of $\Theta$ and there is an $s$-dimensional neighborhood of $\varphi$ that is completely contained in $\Theta$.

2. $\pi_i = f_i(\varphi) > 0$ for $i = \ldots, T$.

3. The mapping $f$ has continuous partial derivatives in a neighborhood of $\varphi$.

4. The Jacobian matrix $(\partial f / \partial \theta)$ is of full rank $s$.

5. The inverse mapping $f^{-1}$ is continuous at $f(\varphi) = \pi$. In other words, for every $\epsilon > 0$ there exists a $\delta > 0$ such that if $||\theta - \varphi|| \geq \epsilon$, then $||f(\theta) - f(\varphi)|| \geq \delta$.

6. The mapping $f$ is continuous at every point $\theta$ in $\Theta$.

Condition 2 ensures that there is no intrinsic empty cell and indeed there are $T$ cells in the multinomial. Condition 3 is a smoothness condition on $f$ and can be replaced by a weaker condition that requires only that $f$ be totally differentiable at $\varphi$, i.e., the first-order linear approximation of $f(\theta)$ at $\varphi$ works as long as the distance between $\theta$ and $\varphi$ approaches zero. Condition 1 ensures that $\Theta$ contains an $s$-dimensional open set around $\varphi$, and Conditions 3 and 4 imply that this open set gets mapped via $f$ into a full $T$-dimensional neighborhood about $\pi$. Conditions 1, 3 and 4 together ensure that the model does have $s$ parameters and not fewer. Condition 5 is a strong identifiability condition on the vector of parameters $\theta$, and it is necessary for global uniqueness of the mapping.
Consistency is a rather weak statistical property of an estimator. Birch [9] proved the efficiency of MLE, or the rate of convergence, under the multinomial setting if the above set of conditions 1–6 are met. Efficiency of an estimator automatically implies the consistency.

**Theorem 2.2.1 (from Bishop, Fienberg, and Holland [11])** Let \( \varphi \) be the true parameter for a parametric model under multinomial sampling, and function \( \mathbf{f} \) maps the parameter into the multinomial cell probability vector. Denote the true cell probability as \( \pi = \mathbf{f}(\varphi) \). Let \( \mathbf{A} \) be the \( T \times s \) matrix whose \((i, j)\) element is \( \pi_i^{-1/2} \left( \frac{\partial f_i(\varphi)}{\partial \theta_j} \right) \). Under Birch’s Conditions 1 – 6, the asymptotic distribution of \( \hat{\theta} \), the MLE for \( \varphi \), as the sample size \( L \to \infty \), is given by

\[
\sqrt{L}(\hat{\theta} - \varphi) \to \mathcal{N}(0, (\mathbf{A}'\mathbf{A})^{-1}) \quad \text{in distribution.}
\]

### 2.2.2 Theoretical Efficiency Results of MLE for Evolutionary Parameters

Recall that we assume that the tree topology and the set of branch lengths are fixed and known. The only unknown parameters that need to be estimated here are the evolutionary parameters in the substitution model.

We now verify that Birch’s conditions hold true for the nucleotide sequence data that follows the F84 substitution model. We note that for a data set with \( N \) sequences that are \( L \) nucleotides long each, the number of multinomial cells is \( 4^N \) (corresponding to \( T \) in the notations when introducing Birch’s conditions in Section 2.2.1), and the sample size is \( L \).

Recall again that the Felsenstein’s five-parameter F84 model has the transition probabilities given by
\[
    P_{ij}(t) = \begin{cases} 
        \pi_j + \pi_j \left( \frac{1}{\Pi_j} - 1 \right) e^{-\mu t} + \left( \frac{\Pi_j - \pi_j}{\Pi_j} \right) e^{-\mu t(K+1)}, & \text{if } i = j; \\
        \pi_j \left( \frac{1}{\Pi_j} - 1 \right) e^{-\mu t} - \left( \frac{\pi_j}{\Pi_j} \right) e^{-\mu t(K+1)}, & \text{if } i \neq j, \text{ transition;} \\
        \pi_j(1 - e^{-\mu t}), & \text{if } i \neq j, \text{ transversion,} 
    \end{cases}
\]

where \( \Pi_j = \pi_A + \pi_G \) if nucleotide \( j \) is a purine (A or G) and \( \Pi_j = \pi_C + \pi_T \) if nucleotide \( j \) is a pyrimidine (C or T). The equilibrium frequencies of the four nucleotides must add to one. As a convention, we designate \( \pi_A, \pi_G, \) and \( \pi_T \) as the three parameters in the model, and it follows that \( \pi_C = 1 - \pi_A - \pi_G - \pi_T \).

The parameter space for \((\mu, K, \pi_A, \pi_G, \pi_T)\) is \((0, \infty) \times (0, \infty) \times \mathcal{S}\), where \((\pi_A, \pi_G, \pi_T)\) is a point from \( \mathcal{S} = \{(a, b, c) : a > 0, b > 0, c > 0, \text{ and } a + b + c < 1\} \). Thus Birch’s Condition 1 in Theorem 2.2.1 is met because the parameter space is an open set in \( \mathbb{R}^5 \). Recall that likelihood at a single site is given by Equation (1.7), and therefore, each multinomial cell probability is in the form of a linear combination of products of terms of \( P_{ij}(t) \)’s given in (2.4). Note that \( P_{ij}(t) \) is positive and is a linear combination of functions of the five parameters that have second-order partial derivatives. Therefore, Conditions 2, 3, and 6 are satisfied. In fact, these four conditions should hold for most general homogeneous Markov substitution models on nucleotide evolution.

Conditions 4 and 5 need more elaborate examinations. The special case for a rooted bifurcating tree with only two external nodes will be established first in Lemma 2.2.1 and Lemma 2.2.2. The general case for a rooted bifurcating tree with \( N \) taxa will be proved in Lemma 2.2.3 and Lemma 2.2.4 by induction.

**Lemma 2.2.1** Let \( T \) represent a rooted bifurcating tree with two external nodes, such that the lengths of the two branches from the root node to child nodes sum to 1. Let \( \mathbf{f} \)
denote the multinomial cell probability vector from a nucleotide evolution that follows the F84 substitution model. Then, the Jacobian matrix \( (\partial \mathbf{f}/\partial \theta) \) is of full rank 5, where \( \theta = (\mu, K, \pi_A, \pi_G, \pi_T) \).

**Proof:** For a bifurcating tree \( T \) with two external nodes, let \( t_1 \) and \( t_2 \) denote the length of the branch connecting the root node and external nodes 1 and 2, respectively. Recalling the time-reversibility property of the substitution model and using the notation for transition probabilities in (2.4), we have, for the multinomial cell probability with nucleotides \( i \) and \( j \) at nodes 1 and 2, respectively,

\[
p_{i,j} = \text{Prob}(X_1 = i, X_2 = j)
= \sum_{k \in \mathcal{A}} \text{Prob}(X_{\text{Root}} = k, X_1 = i, X_2 = j)
= \sum_{k \in \mathcal{A}} \pi_k P_{ki}(t_1) P_{kj}(t_2)
= \sum_{k \in \mathcal{A}} \pi_i P_{ik}(t_1) P_{kj}(t_2)
= \pi_i P_{ij}(t_1 + t_2)
= \pi_i P_{ij}(1),
\]

for any \( i, j \in \mathcal{A} = \{ \text{A, C, G, T} \} \). This is in fact an application of Felsenstein’s pulley principle. To show that the Jacobian matrix is of full rank, it suffices to find the five cell probability functions and verify that the determinant of the Jacobian formed by these five functions is non-zero. Let \( f_1, f_2 \) and \( f_3 \) denote the following three cell probability functions for which the nucleotides at the two external nodes differ by a transversion:

\[
f_1 = p_{A,T} = \pi_A \pi_T (1 - e^{-\mu});
\]

43
\[ f_2 = p_{G,C} = \pi_G (1 - \pi_A - \pi_{G} - \pi_T) (1 - e^{-\mu}); \quad (2.7) \]

and

\[ f_3 = p_{G,T} = \pi_G \pi_T (1 - e^{-\mu}). \quad (2.8) \]

To simplify notations, we also let \( B = 1 - e^{-\mu} \). We select the next two cell probability functions to be those with nucleotides G and A and nucleotides A and A at the external nodes:

\[ f_4 = p_{G,A} = \pi_G \pi_A + \pi_G \pi_A \left( \frac{1}{\pi_A + \pi_G} - 1 \right) e^{-\mu} - \frac{\pi_A \pi_G}{\pi_A + \pi_G} e^{-\mu(K+1)}; \quad (2.9) \]

and

\[ f_5 = p_{A,A} = \pi_A^2 + \pi_A^2 \left( \frac{1}{\pi_A + \pi_G} - 1 \right) e^{-\mu} + \frac{\pi_A \pi_G}{\pi_A + \pi_G} e^{-\mu(K+1)}. \quad (2.10) \]

The partial derivatives of \( f_1 \) are calculated as follows:

\[
\frac{\partial f_1}{\partial \mu} = \pi_A \pi_T e^{-\mu}; \\
\frac{\partial f_1}{\partial K} = 0; \\
\frac{\partial f_1}{\partial \pi_A} = \pi_T B; \\
\frac{\partial f_1}{\partial \pi_T} = \pi_A B; \\
\frac{\partial f_1}{\partial \pi_G} = 0.
\]

Additionally, the partial derivatives of \( f_2 \) and \( f_3 \) are

\[
\frac{\partial f_2}{\partial \mu} = \pi_G (1 - \pi_A - \pi_T - \pi_G) e^{-\mu};
\]

44
\[ \frac{\partial f_2}{\partial K} = 0; \]
\[ \frac{\partial f_2}{\partial \pi_A} = -\pi_G B; \]
\[ \frac{\partial f_2}{\partial \pi_T} = -\pi_G B; \]
\[ \frac{\partial f_2}{\partial \pi_G} = (1 - \pi_A - \pi_T - 2\pi_G)B; \]
\[ \frac{\partial f_3}{\partial \mu} = \pi_G \pi_T e^{-\mu}; \]
\[ \frac{\partial f_3}{\partial K} = 0; \]
\[ \frac{\partial f_3}{\partial \pi_A} = 0; \]
\[ \frac{\partial f_3}{\partial \pi_T} = \pi_G B; \]
\[ \frac{\partial f_3}{\partial \pi_G} = \pi_T B. \]

The partial derivatives of \( f_4 \) and \( f_5 \) are

\[ \frac{\partial f_4}{\partial \mu} = -\pi_A \pi_G \left( \frac{1}{\pi_A + \pi_G} - 1 \right) e^{-\mu} + \frac{\pi_A \pi_G (K + 1)}{\pi_A + \pi_G} e^{-\mu(K + 1)}; \]
\[ \frac{\partial f_4}{\partial K} = \frac{\mu \pi_A \pi_G}{\pi_A + \pi_G} e^{-\mu(K + 1)}; \]
\[ \frac{\partial f_4}{\partial \pi_A} = \pi_G + \left( \frac{\pi_G^2}{(\pi_A + \pi_G)^2} - \pi_A \right) e^{-\mu} - \frac{\pi_G^2}{(\pi_A + \pi_G)^2} e^{-\mu(K + 1)}; \]
\[ \frac{\partial f_4}{\partial \pi_T} = 0; \]
\[ \frac{\partial f_4}{\partial \pi_G} = \pi_A + \left( \frac{\pi_A^2}{(\pi_A + \pi_G)^2} - \pi_A \right) e^{-\mu} - \frac{\pi_A^2}{(\pi_A + \pi_G)^2} e^{-\mu(K + 1)}; \]
\[ \frac{\partial f_5}{\partial \mu} = -\pi_A^2 \left( \frac{1}{\pi_A + \pi_G} - 1 \right) e^{-\mu} - \frac{\pi_A \pi_G (K + 1)}{\pi_A + \pi_G} e^{-\mu(K + 1)}; \]
\[ \frac{\partial f_5}{\partial K} = -\frac{\mu \pi_A \pi_G}{\pi_A + \pi_G} e^{-\mu(K + 1)}; \]
\[ \frac{\partial f_5}{\partial \pi_A} = 2\pi_A + \left( \frac{\pi_A^2 + 2\pi_A \pi_G}{(\pi_A + \pi_G)^2} - 2\pi_A \right) e^{-\mu} + \frac{\pi_A^2}{(\pi_A + \pi_G)^2} e^{-\mu(K + 1)}; \]
\[ \frac{\partial f_5}{\partial \pi_T} = 0; \]
\[ \frac{\partial f_5}{\partial \pi_G} = -\frac{\pi_A^2}{(\pi_A + \pi_G)^2} e^{-\mu} + \frac{\pi_A^2}{(\pi_A + \pi_G)^2} e^{-\mu(K + 1)}. \]
Let \( \mathbf{g} = (f_1, f_2, f_3, f_4, f_5) \), and we next examine the determinant of the Jacobian matrix \( (\partial \mathbf{g}/\partial \theta) \). To simplify the computation, we note that

\[
\det \left( \frac{\partial \mathbf{g}}{\partial \theta} \right) = \det \left( M_1 \frac{\partial \mathbf{g}}{\partial \theta} \right),
\]

where \( M_1 \) is the \( 5 \times 5 \) matrix

\[
\begin{pmatrix}
1 & 0 & 0 & 0 & 0 \\
0 & 1 & 1 & 0 & 0 \\
0 & 0 & 1 & 0 & 0 \\
0 & 0 & 0 & 1 & 0 \\
0 & 0 & 0 & 0 & 1
\end{pmatrix}.
\]

The matrix \( \left( M_1 \frac{\partial \mathbf{g}}{\partial \theta} \right) \) has entries

\[
\begin{pmatrix}
\pi_A \pi_T e^{-\mu} & 0 & \pi_T B & \pi_A B & 0 \\
\pi_G (1 - \pi_A - \pi_G) e^{-\mu} & 0 & -\pi_G B & 0 & (1 - \pi_A - 2\pi_G) B \\
\pi_G \pi_T e^{-\mu} & 0 & 0 & \pi_G B & \pi_T B \\
\frac{\partial f_4}{\partial \mu} & \frac{\partial f_4}{\partial K} & \frac{\partial f_4}{\partial \pi_A} & 0 & \frac{\partial f_4}{\partial \pi_G} \\
\frac{\partial f_5}{\partial \mu} & \frac{\partial f_5}{\partial K} & \frac{\partial f_5}{\partial \pi_A} & 0 & \frac{\partial f_5}{\partial \pi_G}
\end{pmatrix}.
\]

Noting that \( \frac{\partial f_4}{\partial K} = -\frac{\partial f_5}{\partial K} \) and that other entries in the second column equal 0, we have the determinant of the matrix \( \left( M_1 \frac{\partial \mathbf{g}}{\partial \theta} \right) \) in (2.12) is equal to

\[
\begin{vmatrix}
\pi_A \pi_T e^{-\mu} & \pi_T B & \pi_A B & 0 \\
\pi_G (1 - \pi_A - \pi_G) e^{-\mu} & -\pi_G B & 0 & (1 - \pi_A - 2\pi_G) B \\
\pi_G \pi_T e^{-\mu} & 0 & \pi_G B & \pi_T B \\
\frac{\partial (f_4 + f_5)}{\partial \mu} & \frac{\partial (f_4 + f_5)}{\partial \pi_A} & 0 & \frac{\partial (f_4 + f_5)}{\partial \pi_G}
\end{vmatrix}.
\]

46
The determinant of the cominor for the term $\frac{\partial(f_4 + f_5)}{\partial \mu}$ inside the determinant of (2.13) is

$$
\begin{vmatrix}
\pi_T B & \pi_A B \\
-\pi_G B & 0 \\
0 & \pi_G B
\end{vmatrix} = -B^3 \pi_T \pi_G (1 - 2\pi_A - 2\pi_G). \quad (2.14)
$$

Likewise, the determinants of the cominor corresponding to $\frac{\partial(f_4 + f_5)}{\partial \pi_A}$ and $\frac{\partial(f_4 + f_5)}{\partial \pi_G}$ are, respectively,

$$
\begin{vmatrix}
\pi_A \pi_T e^{-\mu} & \pi_A B & 0 \\
\pi_G (1 - \pi_A - \pi_G) e^{-\mu} & 0 & (1 - \pi_A - 2\pi_G) B \\
\pi_G \pi_T e^{-\mu} & \pi_G B & \pi_T B
\end{vmatrix} = -\pi_A \pi_T \pi_G (1 - \pi_A - \pi_G) B^2 e^{-\mu}, \quad (2.15)
$$

and

$$
\begin{vmatrix}
\pi_A \pi_T e^{-\mu} & \pi_T B & \pi_A B \\
\pi_G (1 - \pi_A - \pi_G) e^{-\mu} & -\pi_G B & 0 \\
\pi_G \pi_T e^{-\mu} & 0 & \pi_G B
\end{vmatrix} = -\pi_T \pi_G^2 (1 - \pi_A - \pi_G) B^2 e^{-\mu}. \quad (2.16)
$$

Let $E_1$, $E_2$, and $E_3$ denote the right-hand side expressions of Equations (2.14), (2.15), and (2.16), respectively. It follows that the determinant of the $4 \times 4$ matrix to be multiplied by $\frac{\partial f_4}{\partial K}$ in the expression of (2.13) is equal to

$$
- \frac{\partial(f_4 + f_5)}{\partial \mu} E_1 + \frac{\partial(f_4 + f_5)}{\partial \pi_A} E_2 + \frac{\partial(f_4 + f_5)}{\partial \pi_G} E_3 \\
= -B^3 \pi_T \pi_G (1 - 2\pi_A - 2\pi_G) \pi_A e^{-\mu} (1 - \pi_A - \pi_G) \\
- \pi_A \pi_T \pi_G (1 - \pi_A - \pi_G) B^2 e^{-\mu} \left(2\pi_A + \pi_G + e^{-\mu} (1 - 2\pi_A - \pi_G)\right) \quad (2.17) \\
- \pi_T \pi_G^2 (1 - \pi_A - \pi_G) B^2 e^{-\mu} \left(\pi_A - \pi_A e^{-\mu}\right) \\
= -\pi_A \pi_T \pi_G (1 - \pi_A - \pi_G) B^2 e^{-\mu}.
$$

Combining (2.11), (2.13), and (2.17), we have

47
\[ \det \left( \frac{\partial g}{\partial \theta} \right) = \det \left( M_1 \frac{\partial g}{\partial \theta} \right) \]

\[ \begin{pmatrix}
\frac{\partial f_4}{\partial K} & \frac{\partial f_4}{\partial \mu} & \frac{\partial f_4}{\partial \pi_A} & \frac{\partial f_4}{\partial \pi_T} & \frac{\partial f_4}{\partial \pi_G} \\
\frac{\partial f_5}{\partial K} & \frac{\partial f_5}{\partial \mu} & \frac{\partial f_5}{\partial \pi_A} & \frac{\partial f_5}{\partial \pi_T} & \frac{\partial f_5}{\partial \pi_G} \\
\end{pmatrix} \]

\[ = \frac{\mu \pi_A \pi_T e^{-\mu}}{\pi_A + \pi_T} e^{-\mu(K+1)} \pi_A \pi_T \pi_G(1 - \pi_A - \pi_G) e^{-\mu} \left( 1 - e^{-\mu} \right)^2 \]

< 0.

Lemma 2.2.2 Suppose that \( T \) is a rooted tree with two external nodes that meets the same conditions as in Lemma 2.2.1. Let \( f \) denote the multinomial cell probability vector from the F84 model for the two external nodes of \( T \), then the inverse mapping \( f^{-1} \) is continuous at \( f(\theta) = \pi \) for any \( \theta \) in the parameter space \( \Theta \). In other words, for every \( \epsilon > 0 \) there exists a \( \delta > 0 \) such that if \( ||\theta' - \theta|| \geq \epsilon \), then \( ||f(\theta') - f(\theta)|| \geq \delta \).

Proof: The sketch of proof is as follows. If the distance between the two parameter values \( \theta \) and \( \theta' \) has a lower bound, we examine the cases of when \((\mu, \pi_A, \pi_T, \pi_G)\) and \((\mu', \pi'_A, \pi'_T, \pi'_G)\) are close and when they are not. If they are close, then the distance between \( K \) and \( K' \) has a lower bound, and we can find a multinomial cell probability for which the distance between its values at \( \theta \) and \( \theta' \) is bounded from below above zero. Otherwise, either values of \( \theta \) and \( \theta' \) for a nucleotide frequency parameter differ by a margin larger than a threshold, or all of the frequency parameter values are close and \( \mu \) and \( \mu' \) are apart by more than a certain margin. In either case, we will find sums of cell probabilities, for which the \( L^2 \) metric for the sums has a positive lower

48
bound. Therefore, a lower bound exists for the $L^2$ metric difference of the multinomial cell probability vectors evaluated at $\theta$ and $\theta'$ that are at least $\epsilon$ apart.

We first set up preparation of the main proof. Recall from Equation (2.5) that the probability of the two external nodes with nucleotides $i$ and $j$, $p_{i,j}$, is

$$p_{i,j} = \pi_i P_{i,j}(1).$$

Therefore, the cell probability of having the same nucleotide $A$ at the external nodes of $T$ is

$$g_1 = p_{A,A} = \pi_A^2 + \pi_A^2 \left( \frac{1}{\pi_A + \pi_G} - 1 \right) e^{-\mu} + \frac{\pi_A \pi_G}{\pi_A + \pi_G} e^{-\mu(K+1)}.$$  \hfill (2.18)

In addition, we note that after direct computation, the sums of the following cell probabilities, denoted by $g_2, g_3, g_4,$ and $g_5$, are

$$g_2 = p_{A,T} + p_{A,C} + p_{G,T} + p_{G,C} = (\pi_A + \pi_G)(1 - \pi_A - \pi_G)(1 - e^{-\mu});$$  \hfill (2.19)

$$g_3 = p_{A,A} + p_{A,T} + p_{A,C} + p_{A,G} = \pi_A;$$  \hfill (2.20)

$$g_4 = p_{T,A} + p_{T,T} + p_{T,C} + p_{T,G} = \pi_T;$$  \hfill (2.21)

and

$$g_5 = p_{G,A} + p_{G,T} + p_{G,C} + p_{G,G} = \pi_G.$$  \hfill (2.22)

We note that $g_2$ does not involve the parameter $K$, and $g_3, g_4$ and $g_5$ are simply the parameters of equilibrium frequencies in the model.

Now consider two parameter values $\theta$ and $\theta'$. Denote the four-dimensional vector of components $\mu, \pi_A, \pi_T, \pi_G$ combined for the parameter $\theta$ as $\gamma$, and likewise $\gamma'$ for $\theta'$. In other words, $\theta = (\gamma, K)$ and $\theta' = (\gamma', K')$.

For any $\epsilon > 0$, next we will use the continuity of the function $g_1$ to show that if $\gamma$ and $\gamma'$ are close then the squared difference between $g_1(\gamma, K)$ and $g_1(\gamma', K')$ has a
lower bound provided that $|K - K'| > \frac{\epsilon}{2}$. We start by rewriting the difference of $g_1$ at the two parameter values by adding and subtracting $g_1(\gamma', K)$

\[
g_1(\gamma, K) - g_1(\gamma', K') = g_1(\gamma, K) - g_1(\gamma', K) + g_1(\gamma', K) - g_1(\gamma', K')
\]

(denoted by $D_1$) \hspace{1cm} (denoted by $D_2$). \hspace{1cm} (2.23)

Recall the definition of $g_1$ in (2.18), and we have

\[
D_2 = g_1(\gamma', K) - g_1(\gamma', K') = \frac{\pi_A'\pi_G'}{\pi_A' + \pi_G'} e^{-\mu'(K+1)} \left( 1 - e^{-\mu'(K'-K)} \right).
\] \hspace{1cm} (2.24)

We will denote $D_2$ as $D_2(\gamma', K, K' - K)$ to signify that the dependence of $D_2$ on $K'$ is through $K' - K$. Note that the function $D_2(\gamma', K, K' - K)$ is continuous at $\gamma = (\mu, \pi_A, \pi_T, \pi_G)$ for any fixed $K$ and $K'$. In addition, note that $D_2 > 0$ if $K' - K > 0$ and $D_2 < 0$ if $K' - K < 0$. Consequently, there exists a positive $\eta_1(\theta, \frac{\epsilon}{2})$, such that for $K'$ satisfying $|K' - K| = \frac{\epsilon}{2}$, if

\[
||\gamma - \gamma'||^2 = (\mu - \mu')^2 + (\pi_A - \pi_A')^2 + (\pi_T - \pi_T')^2 + (\pi_G - \pi_G')^2 < \eta_1(\theta, \frac{\epsilon}{2}) \hspace{1cm} (2.25)
\]

the absolute values $|D_2(\gamma', K, K' - K)|$ and $|D_2(\gamma, K, K' - K)|$ (which is positive) are close and

\[
|D_2(\gamma', K, K' - K)| > \frac{1}{2}|D_2(\gamma, K, K' - K)|.
\]
Hence, for \( \gamma' \) that satisfies the condition in (2.25) and for \( K' \) that satisfies \(|K' - K| = \frac{\epsilon}{2}\),

\[
|D_2(\gamma', K, K' - K)| \\
= \left| \frac{\pi' A \pi' G}{\pi_A + \pi_G} e^{-\mu(K + 1)} \left( 1 - e^{-\mu(K' - K)} \right) \right| \\
> \frac{1}{2} \left| \frac{\pi A \pi G}{\pi_A + \pi_G} e^{-\mu(K + 1)} \left( 1 - e^{-\mu(K' - K)} \right) \right| \\
\geq \frac{1}{2} \min \left( \left| \frac{\pi A \pi G}{\pi_A + \pi_G} e^{-\mu(K + 1)} \left( 1 - e^{-\mu|K' - K|} \right) \right|, \left| \frac{\pi A \pi G}{\pi_A + \pi_G} e^{-\mu(K + 1)} \left( e^{\mu|K' - K|} - 1 \right) \right| \right) \\
\Delta = \delta_1(|K' - K|) \quad \text{(by definition)} \\
= \delta_1 \left( \frac{\epsilon}{2} \right). \quad (2.26)
\]

It is useful to note the following properties of the function \( \delta_1 \), defined for \( x > 0 \) by

\[
\delta_1(x) = \frac{1}{2} \min \left( \left| \frac{\pi A \pi G}{\pi_A + \pi_G} e^{-\mu(K + 1)} \left( 1 - e^{-\mu x} \right) \right|, \left| \frac{\pi A \pi G}{\pi_A + \pi_G} e^{-\mu(K + 1)} \left( e^{\mu x} - 1 \right) \right| \right) \quad (2.27)
\]

- \( \delta_1(x) > 0; \)
- \( \delta_1(x) \) depends only on \( \theta \) but not \( \theta'; \)
- \( \delta_1(x) \) is monotonically increasing in \( x > 0. \)

Furthermore, we note that as a function of \( K' - K \), \(|D_2(\gamma', K, K' - K)|\) is increasing in \( K' - K \) when \( K' - K > 0 \), and decreasing in \( K' - K \) when \( K' - K < 0 \). Along with the property of \( \delta_1(x) \) being increasing in \( x \) when \( x > 0 \), we have established that for any \( \epsilon > 0 \), there exists \( \eta_1(\theta, \frac{\epsilon}{2}) \), such that if

\[
||\gamma - \gamma'||^2 = (\mu - \mu')^2 + (\pi_A - \pi'_A)^2 + (\pi_T - \pi'_T)^2 + (\pi_G - \pi'_G)^2 < \eta_1(\theta, \frac{\epsilon}{2}), \quad (2.28)
\]

51
for any \( K' \) that is at least \( \frac{\epsilon}{2} \) apart from \( K \), \( |D_2(\gamma', K, K' - K)| \) has a lower bound, and

\[
|D_2(\gamma', K, K' - K)| = |g_1(\gamma', K) - g_1(\gamma', K')| > \delta_1(\frac{\epsilon}{2}). \tag{2.29}
\]

On the other hand, for the fixed true \( K \), when \( \gamma \) and \( \gamma' \) are sufficiently close, the expression \( D_1 \) is close to 0. In formal terms, there exists \( \eta_2(\theta, \frac{\epsilon}{2}) > 0 \) such that if

\[
||\gamma - \gamma'||^2 = (\mu - \mu')^2 + (\pi_A - \pi'_A)^2 + (\pi_T - \pi'_T)^2 + (\pi_G - \pi'_G)^2 < \eta_2(\theta, \frac{\epsilon}{2}), \tag{2.30}
\]

then

\[
|D_1| < \frac{1}{2} \delta_1(\frac{\epsilon}{2}). \tag{2.31}
\]

Combining the above Inequalities (2.28) through (2.31), we have shown that for a given \( \theta \), there exists \( \eta_3 = \min(\eta_1, \eta_2) > 0 \) that depends only on \( \theta \) and \( \epsilon \), such that if \( ||\gamma - \gamma'||^2 < \eta_3(\theta, \frac{\epsilon}{2}) \), for any \( K' \) satisfying \( |K' - K| \geq \frac{\epsilon}{2} \), the following holds.

\[
(g_1(\gamma, K) - g_1(\gamma', K'))^2 = (g_1(\gamma, K) - g_1(\gamma', K) + g_1(\gamma', K) - g_1(\gamma', K'))^2 \\
\geq (|g_1(\gamma', K) - g_1(\gamma', K')| - |g_1(\gamma, K) - g_1(\gamma', K)|)^2 \\
> \frac{1}{4} \delta_1^2(\frac{\epsilon}{2}). \tag{2.32}
\]

Applying the same derivation to the function \( g_2 \), we can find a lower bound for the squared difference between \( g_2(\gamma) \) and \( g_2(\gamma') \), provided that the difference between \( \pi_A + \pi_G \) and \( \pi'_A + \pi'_G \) is small enough and \( |\mu' - \mu| \) is bounded from below. Formally, for any \( \xi > 0 \), for a given \( \gamma = (\mu, \pi_A, \pi_T, \pi_G) \), there exists \( \eta_4(\gamma, \xi) > 0 \) such that if
\[ |\mu' - \mu|^2 \geq \xi \text{ and } |\pi_A' + \pi_G' - \pi_A - \pi_G|^2 < \eta_4, \]
\[ (g_2(\mu, \pi_A, \pi_T, \pi_G) - g_2(\mu', \pi_A', \pi_T', \pi_G'))^2 > \delta_2^2(\xi), \] (2.33)

where the function \( \delta_2 \) defined by
\[
\delta_2(x) = \frac{1}{4} \min \left( \left| (\pi_A + \pi_G)(1 - \pi_A - \pi_G) e^{-\mu} (1 - e^{-x}) \right|, \right.
\left. \left| (\pi_A + \pi_G)(1 - \pi_A - \pi_G) e^{-\mu} (e^x - 1) \right| \right)
\] (2.34)
is positive and is an increasing function in \( x \) when \( x > 0 \).

Suppose \( \theta = (\mu, K, \pi_A, \pi_T, \pi_G) \) is an arbitrary true parameter in \( \Theta \). Next we will show that for \( \theta \) and an arbitrary \( \epsilon > 0 \), for any \( \theta' \) in \( \Theta \) that is more than \( \epsilon \) apart from \( \theta \), i.e.,
\[ ||\theta' - \theta||^2 = (\mu' - \mu)^2 + (K' - K)^2 + (\pi_A' - \pi_A)^2 + (\pi_T' - \pi_T)^2 + (\pi_G' - \pi_G)^2 \geq \epsilon^2, \] (2.35)

the \( L^2 \) distance between multinomial probability vectors \( \mathbf{f} \) evaluated at \( \theta' \) and \( \theta \) is bounded from below above 0.

With the above preparation, the proof of the lemma is constructed by examining each of the two following cases.

1. If \( \gamma' \) and \( \gamma \) are close, and
\[ ||\gamma' - \gamma||^2 < \min \left( \frac{\epsilon^2}{2}, \eta_3(\theta, \frac{\epsilon}{2}) \right), \] (2.36)
it follows from (2.35) that \( |K' - K| > \frac{\epsilon}{2} \). Hence, combining this with (2.32), we have
\[ ||\mathbf{f}(\theta') - \mathbf{f}(\theta)||^2 \geq (g_1(\gamma, K) - g_1(\gamma', K'))^2 \]
\[ > \frac{1}{4} \delta_1^2(\frac{\epsilon}{2}). \] (2.37)
2. If the condition in \((2.36)\) does not hold, then

\[ ||\gamma' - \gamma||^2 \geq \min \left( \frac{e^2}{2}, \eta_3(\theta, \frac{\epsilon}{2}) \right). \tag{2.38} \]

By replacing \(\xi\) in Inequality \((2.33)\) by \(\frac{1}{4} \min \left( \frac{e^2}{2}, \eta_3(\theta, \frac{\epsilon}{2}) \right)\), it follows that for the given \(\gamma\) there exists a positive \(\eta_4\left(\gamma, \frac{1}{4} \min \left( \frac{e^2}{2}, \eta_3(\theta, \frac{\epsilon}{2}) \right)\right)\), such that if

\[ |\mu' - \mu|^2 \geq \frac{1}{4} \min \left( \frac{e^2}{2}, \eta_3(\theta, \frac{\epsilon}{2}) \right), \tag{2.39} \]

and

\[ |\pi'_A + \pi'_G - \pi_A - \pi_G|^2 < \eta_4, \tag{2.40} \]

then

\[ \left( g_2(\mu, \pi_A, \pi_T, \pi_G) - g_2(\mu', \pi'_A, \pi'_T, \pi'_G) \right)^2 > \delta_2^2 \left( \frac{1}{4} \min \left( \frac{e^2}{2}, \eta_3(\theta, \frac{\epsilon}{2}) \right) \right). \tag{2.41} \]

Consider further the two cases according to whether at least one nucleotide equilibrium frequency parameter value for \(\theta'\) and that for \(\theta\) are not close, and the squared difference is greater than or equal to the threshold

\[ \lambda = \frac{1}{4} \min \left( \min \left( \frac{e^2}{2}, \eta_3(\theta, \frac{\epsilon}{2}) \right), \eta_4 \left( \gamma, \frac{1}{4} \min \left( \frac{e^2}{2}, \eta_3(\theta, \frac{\epsilon}{2}) \right) \right) \right). \tag{2.42} \]

(a) Consider the case in which one of the nucleotide frequency parameters differs by no less than \(\lambda\) for \(\theta'\) and \(\theta\). Without loss of generality, suppose it is the equilibrium frequency of nucleotide A, i.e., the equilibrium frequencies of A in \(\theta'\) and \(\theta\), \(\pi'_A\) and \(\pi_A\), satisfy

\[ (\pi'_A - \pi_A)^2 \geq \lambda. \tag{2.43} \]
Recall from Equation (2.20) that $\pi_A$ can be written as a sum of four cell probabilities, so a positive lower bound can be established for the distance between $f(\theta')$ and $f(\theta)$.

$$
||f(\theta') - f(\theta)||^2 \\
\geq \left( p_{A,A}(\theta') - p_{A,A}(\theta) \right)^2 + \left( p_{A,T}(\theta') - p_{A,T}(\theta) \right)^2 \\
+ \left( p_{A,C}(\theta') - p_{A,C}(\theta) \right)^2 + \left( p_{A,G}(\theta') - p_{A,G}(\theta) \right)^2 \\
\geq \frac{1}{4} \left( p_{A,A}(\theta') + p_{A,T}(\theta') + p_{A,C}(\theta') + p_{A,G}(\theta') - p_{A,A}(\theta) - p_{A,T}(\theta) \\
- p_{A,C}(\theta) - p_{A,G}(\theta) \right)^2 \\
= \frac{1}{4} (\pi'_A - \pi_A)^2 \\
\geq \frac{1}{4} \lambda.
$$

(2.44)

(b) Suppose that the three respective pairs of nucleotide equilibrium frequencies are close, and the squared differences are all less than $\lambda$. Therefore, they are all less than $\frac{1}{4} \min \left( \frac{\epsilon^2}{2}, \eta_3(\theta, \frac{\epsilon}{2}) \right)$. Combined with Inequality (2.38), it leads to

$$
|\mu' - \mu|^2 > \frac{1}{4} \min \left( \frac{\epsilon^2}{2}, \eta_3(\theta, \frac{\epsilon}{2}) \right).
$$

Furthermore, since both

$$
|\pi'_A - \pi_A|^2 < \frac{1}{4} \eta_4 \left( \gamma, \frac{1}{4} \min \left( \frac{\epsilon^2}{2}, \eta_3(\theta, \frac{\epsilon}{2}) \right) \right)
$$

and

$$
|\pi'_G - \pi_G|^2 < \frac{1}{4} \eta_4 \left( \gamma, \frac{1}{4} \min \left( \frac{\epsilon^2}{2}, \eta_3(\theta, \frac{\epsilon}{2}) \right) \right)
$$

hold true, it follows that

$$
|\pi_A + \pi'_G - \pi_A - \pi'_G|^2 < \eta_4 \left( \gamma, \frac{1}{4} \min \left( \frac{\epsilon^2}{2}, \eta_3(\theta, \frac{\epsilon}{2}) \right) \right).
$$
Thus, the conditions for (2.39) and (2.40) are satisfied, and hence,

\[
(g_2(\mu, \pi_A, \pi_T, \pi_G) - g_2(\mu', \pi_A', \pi_T, \pi_G'))^2 > \frac{\delta_2^2}{4} \left( \frac{\epsilon^2}{2}, \eta_3(\theta, \frac{\epsilon}{2}) \right) \cdot
\]

(2.45)

Recall from the definition of \( g_2 \) in Equation (2.19) that \( g_2 \) can be written as a sum of four cell probabilities, and we perform the same manipulation for deriving (2.44) and obtain

\[
\begin{align*}
||f(\theta') - f(\theta)||^2 & \geq \left( p_{A,T}(\theta') - p_{A,T}(\theta) \right)^2 + \left( p_{A,C}(\theta') - p_{A,C}(\theta) \right)^2 \\
& \quad + \left( p_{G,T}(\theta') - p_{G,T}(\theta) \right)^2 + \left( p_{G,C}(\theta') - p_{G,C}(\theta) \right)^2 \\
& \geq \frac{1}{4} \left( p_{A,T}(\theta') + p_{A,C}(\theta') + p_{G,T}(\theta') + p_{G,C}(\theta') - p_{A,T}(\theta) - p_{A,C}(\theta) \\
& \quad - p_{G,T}(\theta) - p_{G,C}(\theta) \right)^2 \\
& = \frac{1}{4} \left( g_2(\gamma') - g_2(\gamma) \right)^2 \\
& > \frac{1}{4} \delta_2^2 \left( \frac{\epsilon^2}{2}, \eta_3(\theta, \frac{\epsilon}{2}) \right) \cdot
\end{align*}
\]

(2.46)

For any \( \theta' \) that satisfies \( ||\theta' - \theta|| \geq \epsilon \), at least one of conditions in Case 1, Case 2(a), and Case 2(b) must be met. Therefore, if we let

\[
\delta = \min \left( \frac{1}{4} \delta_1^2(\frac{\epsilon}{2}), \frac{1}{4} \lambda, \frac{1}{4} \delta_2^2 \left( \frac{1}{4} \min \left( \frac{\epsilon^2}{2}, \eta_3(\theta, \frac{\epsilon}{2}) \right) \right) \right),
\]

it follows from Inequalities (2.37), (2.44) and (2.46) that \( ||f(\theta') - f(\theta)||^2 \geq \delta \) .

\[ \blacksquare \]

**Lemma 2.2.3** Let \( f \) denote the multinomial cell probability vector from a nucleotide evolution that follows the F84 substitution model for a rooted tree, and suppose that the tree topology and branch lengths are known. Then, the Jacobian matrix \( (\partial f/\partial \theta) \) is of full rank 5, where \( \theta = (\mu, K, \pi_A, \pi_G, \pi_T) \).
Proof: We will prove the lemma by induction on the number of external nodes, \( N \).

1. First consider a tree with two nodes. We have proven the proposition in Lemma 2.2.1 for a tree with two nodes for which the lengths of the branches connecting the root node and external nodes add to unity. Suppose that a rooted tree \( T \) with two external nodes has the two branch lengths add up to a constant \( t \). Introduce a variable substitution \( \mu' = \mu \cdot t \), and the cell probabilities of tree \( T \) would correspond to an F84 model with the parameter \( \mu' \) for a tree that has the two branch lengths summing up to 1. Following the same proof in Lemma 2.2.1, we can show the proposition holds true for \( T \).

2. Suppose that the Jacobian matrix is of full rank 5 for any rooted tree with \( n \) external nodes. We now show that the Jacobian matrix is of rank 5 for a rooted tree with \( n + 1 \) external nodes.

For a tree, labeled \( Tree_1 \), with \( n + 1 \) external nodes, where \( n \geq 2 \), there must exist an external node \( b_0 \) whose immediate parent is not the root, as trees considered here are bifurcating trees. Denote the set of all of the external nodes as \( b \). See Figure 2.1 for illustration. Now consider the tree with the selected node \( b_0 \) removed, and label the resulting tree \( Tree_2 \). An observation to note is that for a site pattern in \( Tree_2 \), the multinomial cell probability is equal to the sum of multinomial probabilities of four cells for \( Tree_1 \). These four site patterns for \( Tree_1 \) share the same nucleotide configuration at the set of external nodes \( b \setminus b_0 \), or the entire set of external nodes for \( Tree_2 \), and have nucleotides A, C, G, and T, respectively, at \( b_0 \). This follows from the conditional independence of the Markov process and the total probability principle.
Figure 2.1: Illustration of an \((n+1)\)-external-node rooted tree and the tree with an external node removed. (a) An \((n+1)\)-external-node tree \(Tree_1\). The external node \(b_0\) is not a daughter node of the root. (b) \(Tree_2\), the resulting tree after \(b_0\) is removed from \(Tree_1\).

As for the Jacobian matrix of \(Tree_1\) whose \((i,j)\) element is \(\frac{\partial f_i(\theta)}{\partial \theta_j}\), we claim that the rank is 5. Otherwise, the five columns in the Jacobian matrix for \(Tree_1\) are linearly dependent. In other words, there exist five numbers \(a_1, a_2, a_3, a_4\) and \(a_5\) which are not all equal to 0, such that the resulting linear combination of the partial derivatives for \(f_{Tree_1}\), the vector for the multinomial cell probabilities, is equal to the 0 vector:

\[
a_1 \frac{\partial f_{Tree_1}}{\partial \mu} + a_2 \frac{\partial f_{Tree_1}}{\partial K} + a_3 \frac{\partial f_{Tree_1}}{\partial \pi_A} + a_4 \frac{\partial f_{Tree_1}}{\partial \pi_T} + a_5 \frac{\partial f_{Tree_1}}{\partial \pi_G} = 0. \tag{2.47}
\]

Partition the \(4^{n+1}\) site patterns of \(Tree_1\) into four disjoint sets, each consisting of \(4^n\) site patterns. Each partitioned subset shares a common nucleotide (A, C, G, or T) at the external node \(b_0\). Equation (2.47) implies that the linear
combination restricted on each of the four partitioned subsets of the $4^n$ site patterns, i.e., the vector formed by the selected $4^n$ rows, is equal to 0, a $4^n$-dimensional vector. We can reorder the four subsets of the site patterns such that the $i$-th entry of the resulting vector for each subset has the common nucleotide configuration for $b \setminus b_0$, and is the same as the $i$-th site pattern for Tree$_2$, for any $i = 1, \ldots, 4^n$. Hence, the corresponding four vectors, after possibly rearranging the rows, each a linear combination of partial derivatives for the so selected $4^n$ cell probabilities of Tree$_1$, is 0. Now sum these four $4^n$-dimensional vectors, which share the same coefficients $\{a_1, a_2, a_3, a_4, a_5\} \neq 0$ for the linear combination of partial derivatives. The $i$-th entry of the resulting vector, noting the fact that a cell probability in Tree$_2$ can be decomposed into the sum of the four cell probabilities in Tree$_1$ that share the same nucleotides for nodes $b \setminus b_0$, is equal to the linear combination of the partial derivatives of the $i$-th cell probability of Tree$_2$ with coefficients $\{a_1, a_2, a_3, a_4, a_5\}$, and is also equal to 0. So, that establishes

$$a_1 \frac{\partial f_{\text{Tree}_2}}{\partial \mu} + a_2 \frac{\partial f_{\text{Tree}_2}}{\partial K} + a_3 \frac{\partial f_{\text{Tree}_2}}{\partial \pi_A} + a_4 \frac{\partial f_{\text{Tree}_2}}{\partial \pi_T} + a_5 \frac{\partial f_{\text{Tree}_2}}{\partial \pi_G} = 0. \quad (2.48)$$

This leads to a contradiction with the induction premise. Therefore, the Jacobian for Tree$_1$ with $n + 1$ external nodes is of full rank.

\[\boxed{\text{Lemma 2.2.4}}\hat{\quad} \text{Under the same conditions as in Lemma 2.2.3, the inverse mapping} f^{-1} \text{ is continuous at } f(\varphi) = \pi \text{ for any } \varphi \text{ in the parameter space } \Theta. \text{ In other words, for every } \epsilon > 0 \text{ there exists a } \delta > 0 \text{ such that if } ||\theta - \varphi|| \geq \epsilon, \text{ then } ||f(\theta) - f(\varphi)|| \geq \delta. \]
**Proof:** The proof is constructed by induction on the number of external nodes of the tree as well.

1. We start with a tree with only two external nodes. Lemma 2.2.2 has shown the proposition is true for the special case of a tree with the two branch lengths adding up to 1. Suppose now that a tree $T$ has two external nodes and the two branch lengths add to a constant $t$. Following the same argument used in the proof of Lemma 2.2.3, we introduce the variable substitution $\mu' = \mu \cdot t$. The multinomial probability vector from $T$ with rate parameter $\mu$ is the same as that from a tree $T'$ with two nodes and sum of two branch lengths of 1 under the rate parameter $\mu'$. Applying Lemma 2.2.2 to tree $T'$, it follows that for a true parameter $\varphi' \in \Theta$, $\forall \epsilon > 0$, $\exists \delta > 0$, such that if $||\theta' - \varphi'|| \geq \epsilon$, then

$$||f_{T'}(\theta') - f_{T'}(\varphi')|| \geq \delta,$$

where $\theta' = (\mu', K, \pi_A, \pi_T, \pi_G) = (\mu \cdot t, K, \pi_A, \pi_T, \pi_G)$.

We claim that if $||\theta - \varphi|| \geq \max(\frac{1}{t}, 1) \epsilon$, then

$$||f_T(\theta) - f_T(\varphi)|| \geq \delta,$$

noting that $f_T(\theta) = f_{T'}(\theta')$. This is because for both cases of $t \geq 1$ and $t < 1$ we have $||\theta' - \varphi'|| \geq \epsilon.$
• If \( t \geq 1 \),

\[
\|\theta' - \varphi'\|^2 \\
= \ell^2 (\mu^0 - \mu^0)^2 + (K^\theta - K^\varphi)^2 + (\pi_A^0 - \pi_A^0)^2 + (\pi_T^0 - \pi_T^0)^2 + (\pi_G^0 - \pi_G^0)^2 \\
\geq (\mu^0 - \mu^0)^2 + (K^\theta - K^\varphi)^2 + (\pi_A^0 - \pi_A^0)^2 + (\pi_T^0 - \pi_T^0)^2 + (\pi_G^0 - \pi_G^0)^2 \\
\geq \max\left(\frac{1}{\ell^2}, 1\right) \epsilon^2 \\
\geq \epsilon^2.
\]

• If \( t < 1 \),

\[
\|\theta' - \varphi'\|^2 \\
= \ell^2 (\mu^0 - \mu^0)^2 + (K^\theta - K^\varphi)^2 + (\pi_A^0 - \pi_A^0)^2 + (\pi_T^0 - \pi_T^0)^2 + (\pi_G^0 - \pi_G^0)^2 \\
\geq \ell^2 \left( (\mu^0 - \mu^0)^2 + (K^\theta - K^\varphi)^2 + (\pi_A^0 - \pi_A^0)^2 + (\pi_T^0 - \pi_T^0)^2 + (\pi_G^0 - \pi_G^0)^2 \right) \\
\geq \ell^2 \max\left(\frac{1}{\ell^2}, 1\right) \epsilon^2 \\
= \epsilon^2.
\]

2. Suppose that the proposition holds true for trees with \( n \) external nodes. Now consider a tree with \( n + 1 \) external nodes, labeled as \( Tree_1 \). There exists an external node \( b_0 \) whose immediate parent is not the root node. Remove \( b_0 \) from \( Tree_1 \) and denote the resulting tree as \( Tree_2 \) and it has \( N \) external nodes. Refer to Figure 2.1 for illustration.

Hence, for a true parameter \( \varphi \) and for any \( \epsilon > 0 \), there exists \( \delta > 0 \), such that if another parameter \( \theta \) in the parameter space satisfies

\[
\|\varphi - \theta\| \geq \epsilon,
\]
then

\[ \|f_{\text{Tree}_2}(\theta) - f_{\text{Tree}_2}(\varphi)\| \geq \delta. \]

Note that there are \( 4^n \) multinomial cells for \( f_{\text{Tree}_2} \), and it follows that there

exists a site pattern such that its multinomial probability \( f_{\text{Tree}_2}^i \) evaluated at the

two parameter values \( \theta \) and \( \varphi \) satisfies

\[ \left( f_{\text{Tree}_2}^i(\theta) - f_{\text{Tree}_2}^i(\varphi) \right)^2 \geq \frac{\delta^2}{4^n}. \] (2.49)

Consider the four site patterns for \( \text{Tree}_1 \) that share the same nucleotide con-

figuration at \( b \setminus b_0 \) as the above site pattern for \( \text{Tree}_2 \). These four cell patterns

have nucleotide A, C, G, and T, respectively at node \( b_0 \). Denote the four cor-

responding cell probabilities by \( f_{\text{Tree}_1}^{iA} \), \( f_{\text{Tree}_1}^{iT} \), \( f_{\text{Tree}_1}^{iG} \), and \( f_{\text{Tree}_1}^{iC} \), and we have

\[ f_{\text{Tree}_1}^{iA} + f_{\text{Tree}_1}^{iT} + f_{\text{Tree}_1}^{iG} + f_{\text{Tree}_1}^{iC} = f_{\text{Tree}_2}^i, \] (2.50)

for any parameter value.

The metric distance for \( \text{Tree}_1 \) between multinomial probability vectors for pa-

rameters \( \theta \) and \( \varphi \) that satisfy \( \|\varphi - \theta\| \geq \epsilon \) has a lower bound because

\[
\|f_{\text{Tree}_1}(\theta) - f_{\text{Tree}_1}(\varphi)\|^2 \\
\geq \left( f_{\text{Tree}_1}^{iA}(\theta) - f_{\text{Tree}_1}^{iA}(\varphi) \right)^2 + \left( f_{\text{Tree}_1}^{iT}(\theta) - f_{\text{Tree}_1}^{iT}(\varphi) \right)^2 \\
+ \left( f_{\text{Tree}_1}^{iG}(\theta) - f_{\text{Tree}_1}^{iG}(\varphi) \right)^2 + \left( f_{\text{Tree}_1}^{iC}(\theta) - f_{\text{Tree}_1}^{iC}(\varphi) \right)^2 \\
\geq \frac{1}{4} \left( f_{\text{Tree}_1}^{iA}(\theta) + f_{\text{Tree}_1}^{iT}(\theta) + f_{\text{Tree}_1}^{iG}(\theta) + f_{\text{Tree}_1}^{iC}(\theta) \\
- f_{\text{Tree}_1}^{iA}(\varphi) - f_{\text{Tree}_1}^{iT}(\varphi) - f_{\text{Tree}_1}^{iG}(\varphi) - f_{\text{Tree}_1}^{iC}(\varphi) \right)^2 \\
= \frac{1}{4} \left( f_{\text{Tree}_2}(\theta) - f_{\text{Tree}_2}(\varphi) \right)^2 \\
\geq \frac{\delta^2}{4^{n+1}}. \] (2.51)
Hence, we have shown that $f^{-1}$ is continuous at $f(\varphi)$.

So far, we have verified all six of Birch’s conditions. Hence, under F84 model, we have the following Theorem.

**Theorem 2.2.2** Let $T$ be a bifurcating tree with $N$ external nodes, whose tree topology and branch lengths are fixed and known. Let data $X$ represent the nucleotide sequences at the external nodes, each of which is $L$ nucleotides long. Assume also that the nucleotide evolution follows the F84 model, and denote the true parameter of the model as $\varphi = (\varphi_1, \varphi_2, \varphi_3, \varphi_4, \varphi_5) = (\mu, K, \pi_A, \pi_G, \pi_T)$. Let $\hat{\theta}$ be the MLE of $\varphi$. Then as $L$ approaches infinity, the asymptotic distribution of $\hat{\theta}$ is given by

$$\sqrt{L}(\hat{\theta} - \varphi) \to N(0, (A' A)^{-1}) \quad \text{in distribution},$$

where $A' A$ is the Fisher information matrix, and the $(i, j)$ element is

$$(A' A)_{i,j} = \sum_{k=1}^{4N} \left( \frac{\partial \log(f_k(\varphi))}{\partial \varphi_i} \right) \left( \frac{\partial \log(f_k(\varphi))}{\partial \varphi_j} \right) f_k(\varphi), \quad (2.52)$$

where $f_k(\varphi)$ is the multinomial probability of the $k$-th site pattern under the true parameter $\varphi$.

**Proof:** By the preparation that verifies the six Birch’s conditions, we have established the proposition is true for the rooted trees following Birch’s Theorem 2.2.1 [9]. Noting the time-reversibility property of the substitution model for nucleotide evolution, the likelihood of an unrooted tree is the same as that of the rooted tree with root placed anywhere along the branches. Therefore, the efficiency result is true for unrooted trees as well.
The proof of the Theorem can also apply to the nucleotide evolution that follows the HKY85 model without any difficulty by verifying each of the six Birch's conditions in the same fashion.

None of the commonly used ML tree finding programs really finds the maximum likelihood estimates of the nucleotide equilibrium frequency parameters, $\pi_i$, for $i = A$, G, C or T. These parameters are instead usually estimated by their empirical frequencies in the observed data set, e.g.,

$$\pi_A = \frac{\text{number of nucleotide A in the } N \text{ sequences at external nodes}}{N L},$$

where $N$ is the number of external nodes and $L$ is the number of sites in each sequence.

Several authors (e.g., [38] and [94]) have commented that the empirical frequency estimates are close to the ML estimates. Heuristically, it also fits the intuition as the nucleotides at the external nodes are considered a sample from the stationary distribution. On the other hand, it seems that the contribution of nucleotide frequency parameters to the likelihood as compared to the evolutionary parameters $\mu$ and $K$ is relatively small. The SSA by Salter and Pearl [102] has an option to obtain ML estimates of these frequency parameters, in which the search algorithm assumes other parameters ($\mu$ and $K$) and tree topology and branch lengths vector are all fixed at an estimated value. No software is known to simultaneously find the ML estimate for nucleotide frequency parameters and the $\mu$ and $K$ parameters. Therefore, in the remainder of this Chapter and when discussing estimates of the parameters that are obtained in practice, we will assume the nucleotide frequency parameters are known and fixed, and the variability associated with the MLE of $\mu$ and $K$ does not include the uncertainty of estimation of these nucleotide frequency parameters. However, it is important to note that if a tree searching algorithm simultaneously finds the MLE
of frequency parameters and $\mu$ and $K$, the variance of the limiting distribution of these MLE is given by the inverse of the $5 \times 5$ Fisher information matrix in (2.52) of Theorem 2.2.2, which incorporates the uncertainty for estimating the frequency parameters.

To facilitate the common practice of the exclusion of nucleotide equilibrium frequency parameters when ML estimation is performed in all current phylogenetic analysis packages, the following Theorem is to be used for the asymptotic distribution of the MLE.

**Theorem 2.2.3** Let $T$ be a bifurcating tree with $N$ external nodes, whose tree topology and branch lengths are fixed and known. Let data $X$ represent the nucleotide sequences at the external nodes, each of which is $L$ nucleotides long. Assume also that the nucleotide evolution follows the F84 model with known nucleotide frequency parameters $\pi_A, \pi_G, \pi_T$ and $\pi_C$. Let $\varphi = (\mu, K)$ be the vector of true values of the two parameters. Let $\hat{\varphi}$ denote the MLE of $\varphi$. Then as $L$ approaches infinity, the asymptotic distribution of $\hat{\varphi}$ is given by

$$\sqrt{L}(\hat{\varphi} - \varphi) \Rightarrow N(0, (A'A)^{-1}) \text{ in distribution,}$$

where $A'A$ is the Fisher information matrix

$$\begin{pmatrix}
\sum_{k=1}^{4N} \left( \frac{\partial \log(f_k(\mu, K))}{\partial \mu} \right)^2 f_k(\mu, K) & \sum_{k=1}^{4N} \left( \frac{\partial \log(f_k(\mu, K))}{\partial \mu} \right) \left( \frac{\partial \log(f_k(\mu, K))}{\partial K} \right) f_k(\mu, K) \\
\sum_{k=1}^{4N} \left( \frac{\partial \log(f_k(\mu, K))}{\partial K} \right) \left( \frac{\partial \log(f_k(\mu, K))}{\partial \mu} \right) f_k(\mu, K) & \sum_{k=1}^{4N} \left( \frac{\partial \log(f_k(\mu, K))}{\partial K} \right)^2 f_k(\mu, K)
\end{pmatrix}$$

(2.54)

where $f_k(\mu, K)$ is the multinomial probability of the $k$-th site pattern following the F84 model under the true parameter $\varphi$. 

65
Proof: Birch’s Theorem 2.2.1 [9]. The proof for Theorem 2.2.2 can be applied to the special case of estimation of only two of the parameters as well. □

If all the nucleotide frequency parameters are fixed at 0.25, the above Theorem reduces to the special case of the nucleotide evolution following the K2P model.

2.3 Fisher Information

Although the maximum likelihood approach has been implemented in many tree-building software packages that provide estimates of the evolutionary parameters, only one is known to also provide the variance estimate of the evolutionary parameter estimators. Although the asymptotic efficiency results of the evolutionary parameters had not previously been formally proved, Yang’s PAML [134] provides an estimate of s.e. of the maximum likelihood estimators of the evolutionary parameters via what he terms the curvature method, or the observed Fisher information.

It can been seen from Theorem 2.2.3 that the asymptotic variance-covariance matrix of $\sqrt{L}(\hat{\theta} - \varphi)$ is given by the inverse of Fisher information. Analogous to classical MLE theory for i.i.d. cases, there are two ways to approximate the Fisher information evaluated at the true unknown parameters. One is the observed Fisher information, that is, the Fisher information evaluated at the MLE of parameters given data. In particular, the estimates of variance of the limiting distributions of $\mu$ and $K$ under the conditions of Theorem 2.2.3 are, instead of the sums over all $4^N$ possible site patterns, the sums over the observed $L$ site patterns, possibly not unique $L$ of them, with equal weight given to each site:

$$v\hat{a}_{1}(\mu) = \frac{1}{D} \sum_{k=1}^{L} \left( \frac{\partial \log(f_{k}(\mu, K))}{\partial K} \right)^2$$ (2.55)
and
\[ \hat{v}_{1}(K) = \frac{1}{D} \sum_{k=1}^{L} \left( \frac{\partial \log(f_k(\mu, K))}{\partial \mu} \right)^2, \] (2.56)
where \( f_k(\mu, K) \) is the probability of the site pattern corresponding to the \( k \)-th site in
the matrix representation of data, and \( D \) is \( L^2 \) times the determinant of the matrix
(2.54),
\[ D = \left( \sum_{k=1}^{L} \left( \frac{\partial \log(f_k(\mu, K))}{\partial \mu} \right)^2 \right) \left( \sum_{k=1}^{L} \left( \frac{\partial \log(f_k(\mu, K))}{\partial K} \right)^2 \right) - \left( \sum_{k=1}^{L} \left( \frac{\partial \log(f_k(\mu, K))}{\partial \mu} \right) \left( \frac{\partial \log(f_k(\mu, K))}{\partial K} \right) \right)^2. \] (2.57)
To obtain the variance estimates, we substitute \( \hat{\mu} \) for \( \mu \) and \( \hat{K} \) for \( K \), and note that
the sums in (2.55), (2.56), and (2.57) are taken over the \( L \) observed sites in the data
set. This approach of employing the observed Fisher information is hereafter referred
to as the curvature method or Yang’s method.

The alternative to the observed Fisher information is the expected Fisher information
approach, which involves evaluating the summand of the entries of matrix (2.54)
for each of the \( 4^N \) site patterns at the MLE, summing them up and finding the inverse
of the \( 2 \times 2 \) matrix. To obtain the exact values of the expected Fisher information
evaluated at the maximum likelihood estimators of \( \mu \) and \( K \) requires the evaluations
of and summation over \( 4^N \) terms. As the number of sequences in the data set \( N \) gets
even moderately large, the computational burden is substantial and makes the exact
calculation prohibitive. However, the form of the entries in matrix (2.54) motivates
a Monte Carlo sampling method to approximate the expected Fisher information.

The Monte Carlo sampling (MCS) method works by generating nucleotide \( N \)-
tuples (site patterns) at a large number \( Q \) of sites, or equivalently drawing \( Q \) in-
dependent samples. The nucleotide \( N \)-tuples are randomly drawn according to the
same multinomial distribution that the evolutionary parameters taking values at the MLE, the specified tree and the substitution model determined. In detail, at each of the \( Q \) generated sites, a nucleotide drawn from the stationary distribution of the substitution model is placed at the root node. From top down following the phylogeny, the nucleotide is drawn at the daughter node along each branch according to the \( 4 \times 4 \) edge transition probability matrix determined by the branch length, evolutionary parameters and the nucleotide at the parent node. The above process is repeated until the nucleotides at all \( N \) external nodes are drawn and form an \( N \)-tuple. Next the partial derivatives of the logarithms for the multinomial probability for observing the \( N \)-tuple as the site pattern is evaluated at \((\hat{\mu}, \hat{K})\), and their squares are summed over the \( Q \) generated sites. The variance estimates are thus given by

\[
v\hat{\text{a}r}_2(\hat{\mu}) = \frac{Q}{LD} \sum_{k=1}^{Q} \left( \frac{\partial \log(f_k(\mu, K))}{\partial K} \right)^2 \bigg|_{\mu=\hat{\mu}, K=\hat{K}}
\]

and

\[
v\hat{\text{a}r}_2(\hat{K}) = \frac{Q}{LD} \sum_{k=1}^{Q} \left( \frac{\partial \log(f_k(\mu, K))}{\partial \mu} \right)^2 \bigg|_{\mu=\hat{\mu}, K=\hat{K}},
\]

where \( f_k(\mu, K) \) is the multinomial probability of the site pattern corresponding to the \( k \)-th site in the generated data by MCS, and \( D \) is

\[
D = \left( \sum_{k=1}^{Q} \left( \frac{\partial \log(f_k(\mu, K))}{\partial \mu} \right)^2 \right) \left( \sum_{k=1}^{Q} \left( \frac{\partial \log(f_k(\mu, K))}{\partial K} \right)^2 \right) \\
- \left( \sum_{k=1}^{Q} \left( \frac{\partial \log(f_k(\mu, K))}{\partial \mu} \right) \left( \frac{\partial \log(f_k(\mu, K))}{\partial K} \right) \right)^2 \bigg|_{\mu=\hat{\mu}, K=\hat{K}}.
\]

Both of the MCS and curvature methods are based on a sum over a subset of the \( 4^N \) site patterns, and the difference lies in that the curvature method makes use of only the observed data, whereas MCS employs the simulated data from the \( Q \) (much
greater than \( L \) sites, for which the cell counts are a sample from the multinomial distribution determined by the MLE from the observed data.

It has long been a most controversial issue in the literature as to which provides a better variance estimate between the expected Fisher information and the observed Fisher information. Efron and Hinkley [32] advocated the use of observed information, while some commentators objected to that. The heated debate demonstrated that, in general, one estimator is not universally better than the other. The consensus opinion, nonetheless, seems to suggest that when the likelihood function for a sample with size 1 looks more like a normal density, the observed Fisher information tends to do better; while the expected Fisher information tends to perform well in the other extreme. We will use simulation to compare the performance of the two methods for this particular application of Fisher information in the next section.

2.4 Simulation

Three simulation studies were carried out to evaluate the performance of the MCS method and compare the two variance estimators. Nucleotide data are generated via simulation according to a specified tree. This rooted tree (both the tree topology and branch lengths) was chosen as the ML tree from a real data set of mitochondrial DNA (mtDNA) consisting of 14 species and 231 sites [54] for which the molecular clock assumption seems to hold well. Following the usual convention, the distance between the root node and each external node was fixed at the value of one to make the estimation of the nucleotide substitution rate parameter \( \mu \) possible. The tree topology is shown in Figure 2.2.
Figure 2.2: The ML tree topology inferred from the 14-sequence primate mtDNA set [54]. The branch lengths are not drawn proportional to that of the ML tree.

Four settings of \((\mu, K)\) in the F84 model with all of the nucleotide frequency parameters fixed at 0.25 (or simply the K2P model) were examined in the series of simulations. The values of parameters selected were \(\mu = 0.1\) and \(0.5\) and \(K = 0.5\) and 10.0. If \(K = 0.5\), the probability of observing different nucleotides at two external nodes that diverged from the root node is about 0.175 for \(\mu = 0.1\); while
this probability is about 0.55 for $\mu = 0.5$. If $K = 10$, the probability of observing different nucleotides at two external nodes that diverged from the root node is about 0.49 for $\mu = 0.1$; while this probability is about 0.66 for $\mu = 0.5$. For two external nodes that have coalesced later than the root node, the above probabilities are lower. A value of 0.5 for $K$ represents an expected transition/transversion ratio of 1:1 for a substitution, while a $K$ value of 10 implies that transitions are expected to occur 10.5 times as often as transversions. It is worth mentioning that there are twice as many kinds of transversions as transitions.

In the simulations that were carried out, the program Seq-Gen [95] was employed to generate nucleotide site patterns at the external nodes from a multinomial distribution determined through the specified tree topology, branch lengths vector, evolutionary parameters and the F84 model.

### 2.4.1 Simulation 1

Simulation 1 is intended to assess the Monte Carlo error associated with the different simulation runs generating the same number of sites. Large-sample theory implies that the Monte Carlo error variance for the MCS estimator, based on the generated nucleotide data at $Q$ sites, is of order of $O\left(\frac{1}{Q}\right)$. It remains to be seen from empirical evidence how many sites a simulation has to generate to give a Monte Carlo estimate of expected Fisher information that is reasonably close to the true value.

For the model tree in Figure 2.2 with parameters $(\mu, K)$ set at each of the four combinations, nucleotide data at 40,000 sites were generated in 20 sets of 2,000. Based on the simulated data at each set of 2,000 sites, the variance estimates were obtained
by the MCS method (Equations (2.58), (2.59) and (2.60)). The evaluation of the exact value for the expected Fisher information requires the sum of $4^{14}$ or about $2.7 \times 10^7$ terms. Instead, we appealed to the the MCS estimate based on the the simulated data from the entire 40,000 sites and compared the 20 MCS estimates to the overall estimate. We calculated the square roots of the MCS estimates for the asymptotic variances of MLEs of $\mu$ and $K$, which are the estimates of standard error. Figure 2.3 presents the boxplots for $\hat{s.e.}(\mu)$ divided by the true value of $\mu$ and $\hat{s.e.}(K)$ divided by the true value of $K$. Each of the values that a boxplot summarizes represents an MCS estimate of s.e., divided by the true value of $\mu$ or $K$, based on the simulated data of 2,000 sites in one of the 20 sets. The horizontal dotted line shows the MCS estimate of the s.e., divided by the true value of $\mu$ or $K$, based on the entire 40,000 generated sites. A column in the plot shows the results from one setting of $\mu$ and $K$, and the top panel is for $\mu$, and the bottom panel is for $K$.

The plots suggest that the variability of the MLE for $\mu$, as estimated by MCS method, is similar for the same true $\mu$ value and different levels for true $K$. On the other hand, the variability of MLE for $K$ appears to be largely affected by the true value of $\mu$ as well as the true value of $K$.

We also calculated the ratio of the standard deviation of the 20 estimates to the overall estimate, which is similar to the measure of coefficient of variation with the exception of the overall MCS estimate replacing the average in the denominator. Among the four parameter combinations, the ratios defined above for the s.e. of $\mu$ and $K$ are generally around 2-3%, with the smallest ratio of 1.4% for the s.e. of $\mu$ for $(\mu, K) = (0.5, 10)$, and the biggest ratio of 5.2% for the s.e. of $\mu$ for $(\mu, K) = (0.1, 10)$. 

72
Figure 2.3: Results of Simulation 1. The boxplots for the 20 MCS estimates of s.e. of MLEs for $\mu$ divided by the true value of $\mu$ and those of MLEs for $K$ divided by the true value of $K$ are presented: each MCS estimate is based upon generated nucleotide data from 2,000 sites. The horizontal dotted line represents the overall MCS estimates based on the entire 40,000 sites divided by the true values of $\mu$ and $K$. Note that the limits of all the vertical axes are not the same.
We conclude from the plots and summary of the ratio measure that the Monte Carlo error variability for MCS estimates based on 2,000 generated sites is reasonably small and 2,000 sites will be generated when evaluating MCS estimates in the subsequent simulations for this 14-sequence example. In general, we recommend the number of generated sites to be 10 to 20 times the number of sites in the original data set to obtain an MCS estimate. If the computational cost permits, a larger number of sites can be generated, but the marginal gain is perhaps not substantial.

2.4.2 Simulation 2

This set of simulations is intended to compare the performance of the proposed MCS estimator for the variability of the MLE of the evolutionary parameters $\mu$ and $K$ to that of the observed information approach, under the condition that the true tree topology and the branch lengths are known. Five hundred (500) separate simulation runs were performed for each of the four parameter settings. In each run, a data set of 14 nucleotide sequences with 231 sites was generated. At each site, a nucleotide was drawn with equal probability from $\mathcal{A} = \{A, C, G, T\}$, the stationary distribution for frequency parameters and placed at the root. Following the tree topology from the root node down, a nucleotide at the end of each branch was drawn according to the transition probabilities determined by the branch length, the nucleotide at the top of the branch and the pre-specified values of $\mu$ and $K$. The above process was repeated until nucleotides at all of the external nodes were drawn. This was carried out again by the program Seq-Gen [95]. For each of the 500 generated 231-site 14-sequence data sets, MLEs of $\mu$ and $K$ were searched by a variant of the Newton-Raphson method. The estimates of the variance by the curvature method were computed.
following Equations (2.55), (2.56) and (2.57) with $\mu$ and $K$ replaced by the MLEs. Furthermore, at each pair of estimated $\mu$ and $K$, nucleotide data of 2000 additional sites were generated and MCS estimates of the variances for the MLEs of $\mu$ and $K$ were computed as described in Simulation 1 and in Section 2.3 following Equations (2.58), (2.59) and (2.60).

This simulation produced 500 MLEs of $\mu$ and $K$, each of which was associated with a simulated 231-site 14-sequence nucleotide data set following the F84 substitution model specified by the pre-determined values of $\mu$ and $K$ and the fixed tree topology and branch lengths. The variance and standard deviation based on these 500 estimates were taken as a benchmark for the variability to which estimates from curvature method and from MCS method would be compared.

Figures 2.4 - 2.7 depict the results from Simulation 2. The left and right panels in each Figure represent results for $\mu$ and $K$, respectively. The sub-plots in the first row are the probability histograms for the 500 MLEs of $\mu$ and $K$. They show that the MLEs for $\mu$ and $K$ based on 500 estimates center around the respective true parameter values, and their distributions are approximately normal distributions.

The sub-plots in the second row are the probability histograms for the s.e. estimates from curvature method associated with the MLEs. The sub-plots in the third row are the probability histograms for the s.e. estimates from MCS method associated with the MLEs. The horizontal axes of the corresponding sub-plots in the second row and the third row have the same scale. The vertical lines in the sub-plots in the second and third row represent the standard deviation from the 500 MLEs. The s.e. estimates based on curvature method are underestimates for all but the setting of $(\mu, K) = (0.5, 10)$. The underestimation is most severe when the true
Figure 2.4: Results of Simulation 2, with $\mu = 0.1$ and $K = 0.5$. The left and right panels are for the estimates associated with $\mu$ and $K$, respectively. The sub-plots in the first, second and third row are the probability histograms for the 500 MLEs of $\mu$ and $K$, curvature method s.e. estimates of MLEs, and MCS s.e. estimates of MLEs. The horizontal axes for the corresponding sub-plots in the second and third row have the same scale. The vertical lines represent the standard deviation of the values of the 500 MLEs.
Figure 2.5: Results of Simulation 2, with $\mu = 0.1$ and $K = 10$. The left and right panels are for the estimates associated with $\mu$ and $K$, respectively. The sub-plots in the first, second and third row are the probability histograms for the 500 MLEs of $\mu$ and $K$, curvature method s.e. estimates of MLEs, and MCS s.e. estimates of MLEs. The horizontal axes for the corresponding sub-plots in the second and third row have the same scale. The vertical lines represent the standard deviation of the values of the 500 MLEs.
Figure 2.6: Results of Simulation 2, with \( \mu = 0.5 \) and \( K = 0.5 \). The left and right panels are for the estimates associated with \( \mu \) and \( K \), respectively. The sub-plots in the first, second and third row are the probability histograms for the 500 MLEs of \( \mu \) and \( K \), curvature method s.e. estimates of MLEs, and MCS s.e. estimates of MLEs. The horizontal axes for the corresponding sub-plots in the second and third row have the same scale. The vertical lines represent the standard deviation of the values of the 500 MLEs.
Figure 2.7: Results of Simulation 2, with \( \mu = 0.5 \) and \( K = 10 \). The left and right panels are for the estimates associated with \( \mu \) and \( K \), respectively. The sub-plots in the first, second and third row are the probability histograms for the 500 MLEs of \( \mu \) and \( K \), curvature method s.e. estimates of MLEs, and MCS s.e. estimates of MLEs. The horizontal axes for the corresponding sub-plots in the second and third row have the same scale. The vertical lines represent the standard deviation of the values of the 500 MLEs.
value of $\mu$ is equal to 0.1. Even the largest estimate of s.e. for the parameter setting $(\mu, K) = (0.1, 0.5)$ is less than the standard deviation from the 500 MLEs of $\mu$.

On the other hand, the vertical line representing the s.d. from the 500 MLEs falls in the middle of the histograms for s.e. estimates based on the MCS method. The descriptive statistics (mean and standard deviation) for the s.e. estimates from the two methods, along with the standard deviation of the 500 MLEs are presented in Tables 2.1 and 2.2. The MCS estimate based on the combined 40,000 generated sites in Simulation 1 is presented as well.

<table>
<thead>
<tr>
<th>Simulation 1</th>
<th>Simulation 2</th>
</tr>
</thead>
<tbody>
<tr>
<td>$\mu$</td>
<td>$K$</td>
</tr>
<tr>
<td>0.1</td>
<td>0.5</td>
</tr>
<tr>
<td>0.1</td>
<td>10.0</td>
</tr>
<tr>
<td>0.5</td>
<td>0.5</td>
</tr>
<tr>
<td>0.5</td>
<td>10.0</td>
</tr>
</tbody>
</table>

Table 2.1: Summary of the results for s.e. of $\mu$ in Simulations 1 and 2.

<table>
<thead>
<tr>
<th>Simulation 1</th>
<th>Simulation 2</th>
</tr>
</thead>
<tbody>
<tr>
<td>$\mu$</td>
<td>$K$</td>
</tr>
<tr>
<td>0.1</td>
<td>0.5</td>
</tr>
<tr>
<td>0.1</td>
<td>10.0</td>
</tr>
<tr>
<td>0.5</td>
<td>0.5</td>
</tr>
<tr>
<td>0.5</td>
<td>10.0</td>
</tr>
</tbody>
</table>

Table 2.2: Summary of the results for s.e. of $K$ in Simulations 1 and 2.
The summary statistics on the estimates from the curvature method and the MCS method confirm the findings based on the visual inspection of the plots in Figures 2.4 - 2.7. For the s.e. estimator for MLE of $\mu$, the mean of MCS estimates are very close to the standard deviation of MLE of $\mu$; while curvature method gives estimates much lower than the s.d. except in the parameter setting of $(\mu, K) = (0.5, 10.0)$ where means of s.e. estimates from both methods are equal and also equal to the s.d. The standard deviation of the s.e. estimates for MCS method is either less than or equal to that for curvature method. As for the s.e. estimator for the MLE of $K$, the two methods perform similarly for the parameter $\mu$ value of 0.5. MCS method provides s.e. estimates that are much closer to the s.d., and the estimator from the curvature method shows a downward bias. The standard deviations of the s.e. estimates from the two methods are similar for the true parameter value of $\mu = 0.5$, and that for MCS method is larger for $\mu = 0.1$, in which case curvature method is a seriously biased estimator.

To sum up, Simulation 2 confirms the approximate normal distribution of the MLE of $\mu$ and $K$. Also, the MCS method outperforms the curvature method in providing an estimate for s.e. in most of the situations, and performs at least as well as the curvature method in others. Furthermore, in all likelihood-based tree searching algorithms involving finding the MLE of evolutionary parameters, the implementation of the MCS method adds minimal computational cost when generating the additional sites and evaluating the partial derivatives. In practice, the MCS method is recommended over the curvature method which had been implemented in an available package, PAML [134].
2.4.3 Simulation 3

This set of simulations is intended to assess the performance of the MCS method for estimating the s.e. of MLE of evolutionary parameters under a more realistic scenario, where the tree topology and branch lengths are not known and are required to be estimated. The performance of estimators from the MCS method and the curvature method when the tree topology and branch lengths are simultaneously estimated is evaluated.

Five hundred (500) separate simulation runs were performed for each of the four parameter settings. In each run, a data set of 14 nucleotide sequences with 231 sites was generated by the program Seq-Gen [95] as described for Simulation 2. For each of the 500 generated 231-site 14-sequence data sets, the global MLE for $\mu$ and $K$, and the tree topology and branch lengths was searched by a modification of the SSA method [102]. For each data set, the s.e. estimates for the MLE of $\mu$ and $K$ from the curvature and MCS methods were obtained.

This simulation produced 500 MLEs of $\mu$ and $K$, each of which was associated with a simulated 231-site 14-sequence nucleotide data set while the simultaneous estimation for tree topology and branch lengths was carried out. The variance and standard deviation of these 500 estimates of $\mu$ and $K$ were taken as a benchmark to which the variability estimates from the curvature method and from the MCS method would be compared.

Figures 2.8 – 2.11 depict the results from Simulation 3. Following the same annotation as in Figures 2.4 – 2.7, the left and right panels represents results for $\mu$ and $K$, respectively. The sub-plots in the first row are the probability histograms for the 500 MLEs of $\mu$ and $K$. They show that the MLEs for $\mu$ and $K$ based on 500 estimates
Figure 2.8: Results of Simulation 3, with $\mu = 0.1$ and $K = 0.5$. The left and right panels are for the estimates associated with $\mu$ and $K$, respectively. The sub-plots in the first, second and third row are the probability histograms for the 500 MLEs of $\mu$ and $K$, curvature method s.e. estimates of MLEs, and MCS s.e. estimates of MLEs. The horizontal axes for the corresponding sub-plots in the second and third row have the same scale. The vertical lines represent the standard deviation of the values of the 500 MLEs.
Figure 2.9: Results of Simulation 3 with $\mu = 0.1$ and $K = 10$. The left and right panels are for the estimates associated with $\mu$ and $K$, respectively. The sub-plots in the first, second and third row are the probability histograms for the 500 MLEs of $\mu$ and $K$, curvature method s.e. estimates of MLEs, and MCS s.e. estimates of MLEs. The horizontal axes for the corresponding sub-plots in the second and third row have the same scale. The vertical lines represent the standard deviation of the values of the 500 MLEs.
Figure 2.10: Results of Simulation 3, with $\mu = 0.5$ and $K = 0.5$. The left and right panels are for the estimates associated with $\mu$ and $K$, respectively. The sub-plots in the first, second and third row are the probability histograms for the 500 MLEs of $\mu$ and $K$, curvature method s.e. estimates of MLEs, and MCS s.e. estimates of MLEs. The horizontal axes for the corresponding sub-plots in the second and third row have the same scale. The vertical lines represent the standard deviation of the values of the 500 MLEs.
Figure 2.11: Results of Simulation 3, with $\mu = 0.5$ and $K = 10$. The left and right panels are for the estimates associated with $\mu$ and $K$, respectively. The sub-plots in the first, second and third row are the probability histograms for the 500 MLEs of $\mu$ and $K$, curvature method s.e. estimates of MLEs, and MCS s.e. estimates of MLEs. The horizontal axes for the corresponding sub-plots in the second and third row have the same scale. The vertical lines represent the standard deviation of the values of the 500 MLEs.
center around the respective parameter values, and most of their distributions are approximately normal distributions. The exception is for the distribution of the MLE of $K$ for the parameter setting of $(\mu, K) = (0.5, 10)$, where some large values of the MLE for $K$ were observed. Consistent to the findings in other plots, the distribution of the MLE for $K$ tends to have a long right tail, and it is more pronounced with the higher level of the true parameter value of $K$.

The sub-plots in the second row are the probability histograms for the s.e. estimates from the curvature method associated with the MLEs. The sub-plots in the third row are the probability histograms for the s.e. estimates from the MCS method associated with the MLEs. The horizontal axes of the corresponding sub-plots in the second row and the third row have the same scale. The vertical lines in the sub-plots in the second and third row represent the standard deviation from the 500 MLEs. The s.e. estimates from both methods underestimate the true variability in the MLE, for which a surrogate of the standard deviation for the MLEs is shown by the vertical line. Compared to the curvature method, the MCS method gives s.e. estimates closer to the s.d. The MCS estimator for the s.e. of the MLE for $K$ even provides somewhat fairs estimates in the cases of $(\mu, K) = (0.1, 0.5), (0.1, 10), (0.5, 0.5)$, but that for s.e. of the MLE for $\mu$ is a serious underestimate.

<table>
<thead>
<tr>
<th>$\mu$</th>
<th>$K$</th>
<th>Simulation 2</th>
<th>Simulation 3</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td>SD($\mu$)</td>
<td>SE($\mu$)</td>
</tr>
<tr>
<td>0.1</td>
<td>0.5</td>
<td>0.0109</td>
<td>0.0087</td>
</tr>
<tr>
<td>0.1</td>
<td>10.0</td>
<td>0.0109</td>
<td>0.0083</td>
</tr>
<tr>
<td>0.5</td>
<td>0.5</td>
<td>0.0290</td>
<td>0.0268</td>
</tr>
<tr>
<td>0.5</td>
<td>10.0</td>
<td>0.0289</td>
<td>0.0289</td>
</tr>
</tbody>
</table>

Table 2.3: Summary of the results for s.e. of $\mu$ in Simulations 2 and 3.
Table 2.4: Summary of the results for s.e. of $K$ in Simulations 2 and 3.

The descriptive statistics (mean and standard deviation) for the s.e. estimates from the two methods, along with the standard deviation of the MLEs are presented in Tables 2.3 and 2.4. The standard deviation of MLEs from Simulation 2, as well as the mean of s.e. estimates from the two methods, are presented as a reference.

Compared to the s.e. estimates in Simulation 2, both methods give similar s.e. estimates in Simulation 3. Even though the MCS method gives larger s.e. estimates than the curvature method, the estimates are markedly less than the standard deviation of MLEs. The reason is that the MCS method does not take into account the variability that the simultaneous estimation of tree topology and branch lengths introduces to the estimation of the evolutionary parameters, and causes the systematic downward bias. This is also demonstrated by the larger standard deviation of MLEs in Simulation 3, where the tree topology and branch lengths were simultaneously estimated, as compared to the standard deviation in Simulation 2. In practice, when the tree topology and branch lengths are required to be estimated, the variability of the evolutionary parameter MLE needs to be estimated by alternative approaches.
CHAPTER 3

LIMITING DISTRIBUTION OF MLE WHEN TREE IS UNKNOWN

In the last Chapter, we have shown the asymptotic efficiency results for the MLE of evolutionary parameters when tree topology and branch lengths are known. Also, the MCS method was demonstrated to provide a good estimate for the s.e. of MLE of evolutionary parameters when the tree is known, and it performs better than the method of observed Fisher information that had been the only estimator in the literature. In the case that the tree is unknown and needs to be estimated, the impact is assessed of using the variance estimator from the Fisher information, which ignores the uncertainty associated with the estimation of tree. If the impact is relatively small, one had hoped that it might be used as a practical rough estimator. Simulation 3 results reveal that the approach does not satisfactorily estimate the true variability when the tree is unknown and needs to be estimated and the underestimation is severe.

In this Chapter, the large-sample property of the joint MLE of tree topology, branch lengths and evolutionary parameter is established for a typical substitution model, namely the five-parameter F84 or HKY85 model. We will demonstrate that, when the joint MLE of tree topology, branch lengths, and evolutionary parameters
is found, the tree topology estimator converges to the true topology as the number of sites approaches $\infty$. Moreover, it will be proved that the MLE for branch lengths and evolutionary parameters are asymptotically efficient.

A practical algorithm for the variance estimator of the MLE for evolutionary parameters will be proposed in the next Chapter.

3.1 Limiting Distribution of Joint MLE when the Tree Topology is Known

At first, consider the case of the tree topology being fixed at the true topology, or assuming that the tree topology is known. We will establish that the asymptotic efficiency of the joint MLE of tree branch lengths and evolutionary parameters by verifying Birch’s conditions.

We consider unrooted trees. Although a rooted tree has the same likelihood as the corresponding unrooted tree, the location of the root cannot be determined just based on the data at external nodes alone without invoking the assumption of molecular clock or having an outgroup. Therefore, only unrooted tree can be constructed purely based on the ML method. We also restrict the tree space to that for bifurcating trees only, that is, each internal node has three and only three nodes (internal or external) connected to it, or each internal node is of degree three. A star phylogeny is not allowed.

Let $N$ denote the number of external nodes, and $T$ a known true tree topology. Let $\mathbf{t}$ denote the vector of branch lengths for $T$. The vector $\mathbf{t}$ can be characterized as the vector of lengths for edges along $T$, which also satisfies the identifiability requirement with the overall substitution rate parameter $\mu$. There are many equivalent parameterizations on $\mathbf{t}$ and $\mu$ combined. For example, one is to set a specified branch to
have unit length and to treat $\mu$ as a parameter. We once again consider Felsenstein’s five-parameter F84 model, for which the transition probabilities are shown in (2.4). Let $\tau$ denote the evolutionary parameter: $\tau = (\mu, K, \pi_A, \pi_G, \pi_T)$. The parameter of the unknowns to be estimated are all interval-scaled. Denote it by $\theta$, and we have $\theta = (t, \tau)$. Let $\Theta$ be the set of all possible values of $\theta$ such that the branch lengths vector $\mathbf{t}$ and $\tau$ are identifiable and $\mathbf{t}$ is congruent with the specified tree topology $T$.

Recall Birch’s conditions in Section 2.2.1, and these conditions will be verified to be true, under the above setting, as follows.

1. The true parameter value is an interior point of $\Theta$. This is because $T$ is restricted to be a bifurcating tree, and the branch lengths vector can take on values from only an open set. So is the case for the evolutionary parameters, using the same argument in proving Theorem 2.2.2.

2. Each nucleotide configuration for the set of external nodes has a positive probability associated, as the probability can be written as a sum of products of linear combinations of exponential terms as in Equation (1.7), each of which is positive. Note that Equation (1.7) can be easily adapted to an unrooted tree, as the likelihood for a rooted tree and its corresponding unrooted tree is identical.

3. The mapping $\mathbf{f}$, the cell probability vector as a function of $\theta$, has continuous partial derivatives in a neighborhood of the true parameter, because of the form of the cell probabilities as sum of terms in Equation (1.7).

4. The Jacobian matrix $(\partial \mathbf{f}/\partial \theta)$ is of full rank. This will be proved in an ensuing lemma.
5. The inverse mapping \( f^{-1} \) is continuous at \( f(\theta) \), the multinomial cell probability vector for the true parameter. This will be established in an ensuing lemma.

6. The mapping \( f \) is continuous at every point \( \theta \) in \( \Theta \), which is verified again from the form of cell probabilities.

To set up the proofs of the lemmas for Conditions 4 and 5, we note the following proposition.

**Proposition 3.1**

Let \( T_1 \) denote an unrooted tree with \( N \) \((N \geq 3)\) external nodes, and \( v_0 \) be an external node. Let \( T_2 \) denote the resulting tree with \( v_0 \) removed from \( T_1 \). By definition of an unrooted bifurcating tree, each internal node has degree 3, that is, three nodes are connected to an internal node through the edges in the bifurcating tree. Denote the internal node of \( T_1 \) that is directly linked to \( v_0 \) by \( v_3 \), and the other two nodes linked to \( v_3 \) by \( v_1 \) and \( v_2 \). Note that these two nodes \( v_1 \) and \( v_2 \) could be both internal nodes, both external nodes (when \( N = 3 \)) or one of each. There must exist an external node \( b_1 \) (it is the same as \( v_1 \) if \( v_1 \) is an external node), such that the path from \( v_0 \) to \( b_1 \) along the topology of \( T_1 \) contains the edges from \( v_0 \) to \( v_1 \) through \( v_3 \). Similarly, there must exist an external node \( b_2 \), possibly same as \( v_2 \), such that the path from \( v_0 \) to \( b_2 \) contains the edges from \( v_0 \) to \( v_2 \) through \( v_3 \).

The sets of edges in trees \( T_1 \) and \( T_2 \) only differ at the local neighborhood of node \( v_3 \). If we compare the branch lengths vector for \( T_1 \) and that for \( T_2 \), the difference is that the vector for \( T_2 \) has an entry of the length for the edge connecting the two adjacent nodes \( v_1 \) and \( v_2 \), \( t_{v_1 v_2} \). In its place, the vector for \( T_1 \) has, in the neighborhood of \( v_3 \), three entries – the lengths of three edges \( t_{v_1 v_3} \), \( t_{v_2 v_3} \), and \( t_{v_0 v_3} \), while the rest of the branch lengths vector outside of the path between \( v_1 \) and \( v_2 \) are the same.
for the two trees. In tree $T_1$, let $d_{v_0 b_1}$ represent the sum of lengths of edges for the unique path connecting $v_0$ and $b_1$, and $d_{v_0 b_2}$ the sum of lengths of edges for the path connecting $v_0$ and $b_2$. Denote the vector of branch lengths for $T_2$ with the entry $t_{v_1 v_2}$ removed as $\mathcal{B}$. The symbol $\parallel$ denotes the concatenation of two vectors. Given the topology of $T_1$ and the set of branch lengths $\mathcal{B}$, the vector $(t_{v_1 v_3}, t_{v_2 v_3}, t_{v_0 v_3})$ uniquely determines the vector $(t_{v_1 v_2}, d_{v_0 b_1}, d_{v_0 b_2})$, as each of the three entries in the latter vector is the addition of the edge lengths in $T_1$, or the entries of the vector $\mathcal{B} \parallel (t_{v_1 v_3}, t_{v_2 v_3}, t_{v_0 v_3})$. Conversely, if a vector value of $(t_{v_1 v_2}, d_{v_0 b_1}, d_{v_0 b_2})$ is congruent with a tree with topology $T_1$, then it uniquely determines $(t_{v_1 v_3}, t_{v_2 v_3}, t_{v_0 v_3})$. The open set of the possible values of $\mathcal{B} \parallel (t_{v_1 v_3}, t_{v_2 v_3}, t_{v_0 v_3})$ in the Euclidean space $\mathbb{R}^{2(N-2)}$ is mapped to an open set for $\mathcal{B} \parallel (t_{v_1 v_2}, d_{v_0 b_1}, d_{v_0 b_2})$ in $\mathbb{R}^{2(N-2)}$. Moreover, the metrics
for the two open sets are equivalent, that is, two points in one space are close if and only if their mapped points are close. In formal terms, the mapping between the two parameterizations is a linear transformation,

\[ \mathcal{B} \| (t_{v_1v_2, d_{v_0b_1}, d_{v_0b_2}}) = \mathcal{B} \| (t_{v_1v_3, t_{v_2v_3}, t_{v_0v_3}}) M^T, \]

where \( M \) is a \( 2(N - 2) \times 2(N - 2) \) matrix

\[
\begin{pmatrix}
I_{2N-7} & 0 & 0 & 0 \\
0 & 1 & 1 & 0 \\
u_1 & 1 & 0 & 1 \\
u_2 & 0 & 1 & 1
\end{pmatrix}_{(2N-4) \times (2N-4)}
\]

whose first \( 2N - 7 \) rows form \((I_{2N-7} \quad 0)\). The \( 2N - 6^{th} \) row consists of the values of 1 for the diagonal entry and the next entry to its right and 0’s for the remaining entries. The first \( 2N - 7 \) entries of last two rows are vector \( u_1 \) and \( u_2 \), consisting of entries of either 1 or 0. Moreover, the matrix \( M \) is of full rank. Denote the largest eigenvalue of \( M \) by \( J \ ( J \geq 1 ) \), then for two points in the space for \( \mathcal{B} \| (t_{v_1v_3}, t_{v_2v_3}, t_{v_0v_3}) \) with distance \( d \), their mapped points in the space of \( \mathcal{B} \| (t_{v_1v_2, d_{v_0b_1}, d_{v_0b_2}}) \) have distance between \( d/J \) and \( d \cdot J \). This can be easily verified by the canonical decomposition of the matrix \( M \) which is of full rank.

Lemma 3.1.1 Let \( T \) denote the topology of an unrooted tree with \( N \) external nodes with \( N \geq 3 \). Let \( t \) denote the branch lengths vector in \( \mathbb{R}^{2(N-2)} \). Let \( \tau = (\mu, K, \pi_A, \pi_G, \pi_T) \) represent the evolutionary parameters in the F84 model. Let \( f \) denote the multinomial cell probability vector that follows the nucleotide evolution of the F84 model with the fixed tree topology \( T \), and \( \theta = (\tau, t) \) as parameter of branch lengths and evolutionary parameters, then the Jacobian matrix \((\partial f/\partial \theta)\) is of full rank.
Proof: We will prove the lemma by induction on the number of external nodes, $N$.

1. First we will prove the case for an unrooted tree, denoted $T_1$ (Figure 3.2 (a)), with 3 external nodes, nodes 1, 2 and 3. The internal node is labeled node 4. Due to the identifiability requirement of $\mu$ and the branch lengths vector, there are only two free branch lengths variables. Without loss of generality, we can assume the parameterization is such that the length of the path connecting two external nodes 1 and 2 is 1, and the branch lengths vector is $(t_{14}, t_{34})$. The goal is to show that $(\partial \mathbf{f} / \partial \theta)$ is of rank 7 for $\theta = (\tau, \mathbf{t}) = (\mu, K, \pi_A, \pi_G, \pi_T, t_{14}, t_{34})$.

![Diagram of trees](https://via.placeholder.com/150)

(a) Tree $T_1$ with 3 external nodes  
(b) Tree $T_2$ with only nodes 1 and 2 after node 3 is removed

Figure 3.2: Illustration of a three-external-node unrooted tree and the tree with an external node removed. (a) An three-external-node tree $T_1$. The external node 3 is any external node. The edge lengths add to one for the edge between nodes 1 and 4 and between nodes 2 and 4. (b) $T_2$, the resulting tree after the branch between nodes 3 and 4 is removed from $T_1$. It contains just two external nodes, and the length of the only branch is one.
Consider the tree $T_2$ (Figure 3.2 (b)), formed by removing node 3 from $T_1$, which is now composed of external nodes 1 and 2. As was shown in Lemma 2.2.1, noting that the likelihood of a rooted tree is equal to that of the corresponding unrooted tree, there exist five nucleotide configurations at nodes 1 and 2 for tree $T_2$ such that for the multinomial probabilities of observing them, $g_1, g_2, g_3, g_4,$ and $g_5$, the Jacobian matrix with respect to $\tau$ satisfies that

$$\det \left( \frac{\partial g_1}{\partial \tau}, \frac{\partial g_2}{\partial \tau}, \frac{\partial g_3}{\partial \tau}, \frac{\partial g_4}{\partial \tau}, \frac{\partial g_5}{\partial \tau} \right) \neq 0. \quad (3.3)$$

We note that each of the above five probabilities, $g_i$, is a sum of four multinomial probabilities of $\mathbf{f}$ for the three-external-node tree $T_1$. These four nucleotide configurations for $T_1$ have the nodes 1 and 2 take on the same nucleotide configuration as $g_i$ and node 3 take on nucleotide A, C, G and T, respectively. This follows from the total probability principle.

Consider the probabilities of observing the following nucleotides at two of the external nodes for $T_1$:

$$g_6 = \text{Prob(Node 1 has nucleotide A, Node 3 has nucleotide T)}$$

$$= \pi_A \pi_T (1 - e^{-\mu(t_{14} + b_{44})}); \quad (3.4)$$

and

$$g_7 = \text{Prob(Node 2 has nucleotide A, Node 3 has nucleotide T)}$$

$$= \pi_A \pi_T (1 - e^{-\mu(1-t_{14}+t_{44})}), \quad (3.5)$$
for which the calculations of these probabilities follow the argument for deriving (2.5). Once again, each of $g_6$ and $g_7$ is a sum of four multinomial cell probabilities in vector $\mathbf{f}$ for the three-external-node tree $T_1$.

By direct calculation, we have

\[
\frac{\partial g_6}{\partial t_{14}} = \pi_A \pi_T \mu e^{-\mu (t_{14} + t_{34})};
\]
\[
\frac{\partial g_6}{\partial t_{34}} = \pi_A \pi_T \mu e^{-\mu (t_{14} + t_{34})};
\]
\[
\frac{\partial g_7}{\partial t_{14}} = -\pi_A \pi_T \mu e^{-\mu (1-t_{14} + t_{34})};
\]

and

\[
\frac{\partial g_7}{\partial t_{34}} = \pi_A \pi_T \mu e^{-\mu (1-t_{14} + t_{34})}.
\]

Hence, the determinant of the associated $2 \times 2$ matrix is

\[
\begin{vmatrix}
\frac{\partial g_6}{\partial t_{14}} & \frac{\partial g_7}{\partial t_{14}} \\
\frac{\partial g_6}{\partial t_{34}} & \frac{\partial g_7}{\partial t_{34}} \\
\end{vmatrix} = 2(\pi_A \pi_T \mu)^2 e^{-\mu(1+2t_{34})} > 0. \tag{3.6}
\]

Since $g_i$, $i = 1, \cdots, 5$, do not involve $t_{14}$ or $t_{34}$, their partial derivatives with respect to a branch length variable are all 0. It follows that the Jacobian determinant of $\mathbf{g} = (g_1, g_2, g_3, g_4, g_5, g_6, g_7)$ with respect to $\theta = (\tau, \mathbf{t}) = (\mu, K, \pi_A, \pi_G, \pi_T, t_{14}, t_{34})$ is
\[
\begin{align*}
\det \left( \frac{\partial g}{\partial \theta} \right) &= \\
&= \det \left( \frac{\partial g_1}{\partial \tau}, \frac{\partial g_2}{\partial \tau}, \frac{\partial g_3}{\partial \tau}, \frac{\partial g_4}{\partial \tau}, \frac{\partial g_5}{\partial \tau}, \frac{\partial g_6}{\partial t_{14}}, \frac{\partial g_7}{\partial t_{14}} \right) \\
&\quad \left| \begin{array}{cccc}
\frac{\partial g_6}{\partial t_{14}} & \frac{\partial g_7}{\partial t_{14}} \\
\frac{\partial g_6}{\partial t_{34}} & \frac{\partial g_7}{\partial t_{34}} \\
\end{array} \right|_{7 \times 7} \\
\neq 0.
\end{align*}
\]

Thus, we have shown that there exist seven linear combinations of multinomial cell probabilities among \( f \), so that their Jacobian matrix is of rank 7. Therefore, \( \partial f/\partial \theta \) is of rank 7.

2. Suppose that the Jacobian is of full rank for any unrooted tree with \( n \) \((n \geq 3)\) external nodes. Now consider an unrooted tree \( T_1 \) with \( n + 1 \) external nodes. Refer to Figure 3.1 for illustration. Let \( v_0 \) be an external node, and \( T_2 \) denote the tree with \( v_0 \) removed from \( T_1 \). \( \mathcal{B} \) denotes the branch lengths vector for \( T_2 \) with \( t_{v_1v_2} \) excluded. It follows from Proposition 3.1 that there exist two external nodes \( b_1 \) and \( b_2 \), and that two parameterizations of branch lengths vector for \( T_1 \) differ by a linear transformation \( \mathcal{M} \), a \( 2(n - 1) \times 2(n - 1) \) matrix, that is of full rank.

Upon the induction premise, for tree \( T_2 \), there exist \( 2n + 1 \) nucleotide configurations so that the Jacobian of the probabilities of observing these configurations at the external nodes of \( T_2 \), namely \( g_1, \ldots, g_{2n+1} \), is of full rank with respect to the parameter vector, \( \theta_{T_2} \), consisting of \( \tau = (\mu, K, \pi_A, \pi_G, \pi_T) \) and the \((2n-4)\)-dimensional branch lengths vector \( \mathcal{B} \| t_{v_1v_2} \). Note that each of these probabilities,
\(g_i\), is a sum of four multinomial cell probabilities among \(f\) for \(T_i\). The partial
derivatives of \(g_i\)'s with respect to both \(d_{\nu_0b_1}\) and \(d_{\nu_0b_2}\) are 0.

Similar to that in the proof for the case of three external nodes, consider the
following two probabilities:

\[
g_{2n+2} = \text{Prob} (\text{Node } \nu_0 \text{ has nucleotide A, Node } b_1 \text{ has nucleotide T})
\]
\[
= \pi_A \pi_T (1 - e^{-\mu d_{\nu_0b_1}}); \quad \text{(3.8)}
\]

and

\[
g_{2n+3} = \text{Prob} (\text{Node } \nu_0 \text{ has nucleotide A, Node } b_2 \text{ has nucleotide T})
\]
\[
= \pi_A \pi_T (1 - e^{-\mu d_{\nu_0b_2}}). \quad \text{(3.9)}
\]

We note that \(g_{2n+2}\) is the sum of \(4^{n-1}\) multinomial cell probabilities in \(f\) that
share the same nucleotides A and T at nodes \(\nu_0\) and \(b_1\), and \(g_{2n+3}\) is the sum of
\(4^{n-1}\) multinomial cell probabilities in \(f\) that share the same nucleotides A and
T at nodes \(\nu_0\) and \(b_2\).

Direct computation yields that the determinant of the following matrix

\[
\begin{vmatrix}
\frac{\partial g_{2n+2}}{\partial d_{\nu_0b_1}} & \frac{\partial g_{2n+3}}{\partial d_{\nu_0b_1}} \\
\frac{\partial g_{2n+2}}{\partial d_{\nu_0b_2}} & \frac{\partial g_{2n+3}}{\partial d_{\nu_0b_2}}
\end{vmatrix} = \left(\pi_A \pi_T \mu\right)^2 e^{-\mu (d_{\nu_0b_1} + d_{\nu_0b_2})} > 0. \quad \text{(3.10)}
\]

Following the same derivation as in the formulation for \(3.7\), for \(g = (g_1, \ldots, g_{2n+3})\)
and \(\xi = \tau \|B\| (t_{0v_2}, d_{\nu_0b_1}, d_{\nu_0b_2})\), we have

\[
\det \left( \frac{\partial g}{\partial \xi} \right) \neq 0. \quad \text{(3.11)}
\]
The parameterizations of $\xi$ and $\theta = \tau \|B\|(t_{v_1 v_2}, t_{v_2 v_3}, t_{v_3 v_4})$ differ by a linear transformation $\begin{pmatrix} I_5 & 0 \\ 0 & M \end{pmatrix}$ following (3.1) in Proposition 3.1, and therefore,

$$\det \left( \frac{\partial g}{\partial \theta} \right) = \det \left( \frac{\partial g}{\partial \xi} \right) \begin{vmatrix} I_5 & 0 \\ 0 & M \end{vmatrix} \neq 0. \quad (3.12)$$

Thus, there exist $2n + 3$ linear combinations of multinomial cell probabilities in $f$ so that their Jacobian matrix is of full rank. It follows that $(\partial f/\partial \theta)$ is of full rank.

\[\square\]

**Lemma 3.1.2** Under the same conditions as in Lemma 3.1.1, the inverse mapping $f^{-1}$ is continuous at $f(\theta)$, the multinomial cell probability for the true parameter. Equivalently, for every $\epsilon > 0$ there exists a $\delta > 0$ such that if $||\theta' - \theta|| \geq \epsilon$, then $||f(\theta') - f(\theta)|| \geq \delta$.

**Proof:** Again, we prove the lemma by induction on the number of external nodes, $N$.

1. Consider an unrooted tree, $T_1$, with three external nodes, 1, 2 and 3 (Figure 3.2).

The parameter $\theta$ is the vector of $\tau ||(t_{14}, t_{34})$ assuming the same parameterization of the tree as in the proof of the case of three external nodes in Lemma 3.1.1 and that the sum of edge lengths between nodes 1 and 2 is one. Let $\xi$ denote the vector of $\tau ||(d_{13}, d_{23})$, where $d_{13} = t_{14} + t_{43}$ and $d_{23} = 1 - t_{14} + t_{43}$. By Proposition 3.1, the Euclidean metric on $\xi$ is equivalent to that on $\theta$. Therefore, it suffices to show that for every $\epsilon > 0$ there exists a $\delta > 0$ such that if $||\xi' - \xi|| \geq \epsilon$, then $||f(\xi') - f(\xi)|| \geq \delta$. 

100
Consider the tree $T_2$, formed by removing node 3 from $T_1$, which comprises the two external nodes 1 and 2. The multinomial cell probability vector of tree $T_2$ is denoted as $\mathbf{g}$. Recall Lemma 2.2.2, and we have for every $\epsilon_2 > 0$ there exists a $\delta_2(\epsilon_2) > 0$ (that may depend on $\tau$) such that if $||\tau' - \tau|| \geq \epsilon_2$, then $||\mathbf{g}(\tau') - \mathbf{g}(\tau)|| \geq \delta_2(\epsilon_2)$. We note again that each cell probability of $\mathbf{g}$ is a sum of four probabilities in $\mathbf{f}$ for tree $T_1$, and it follows that

$$||\mathbf{f}(\xi') - \mathbf{f}(\xi)||^2 \geq \frac{1}{4}||\mathbf{g}(\tau') - \mathbf{g}(\tau)||^2 \geq \frac{1}{4}\delta_2^2(\epsilon_2). \quad (3.13)$$

Consider the function $h$ of $x$, defined for $\tau = (\mu, K, \pi_A, \pi_G, \pi_T)$ as

$$h_{\tau}(x) = \pi_A\pi_T(1 - e^{-\mu x}). \quad (3.14)$$

The following two probabilities $h_1$ and $h_2$ for tree $T_1$ are computed:

$$h_1 = \text{Prob}(\text{Node 1 has nucleotide A, Node 3 has nucleotide T})$$
$$= \pi_A\pi_T(1 - e^{-\mu d_{13}})$$
$$= h_{\tau}(d_{13}); \quad (3.15)$$

and

$$h_2 = \text{Prob}(\text{Node 2 has nucleotide A, Node 3 has nucleotide T})$$
$$= \pi_A\pi_T(1 - e^{-\mu d_{23}})$$
$$= h_{\tau}(d_{23}). \quad (3.16)$$

Following the same derivation in (2.23) through (2.32) for the proof of Lemma 2.2.2, we can show that the squared distance between $h_{\tau}(x)$ and $h_{\tau}(x')$ has a
lower bound, provided that the distance between $\tau$ and $\tau'$ is small enough and $|x' - x|$ is bounded from below. Formally, for any $\epsilon_1 > 0$, for any given $\tau$ and $x$, there exists $\eta > 0$ that depends only on $\tau$, $x$, and $\epsilon_1$, such that if $|x' - x| > \epsilon_1$ and $\|\tau' - \tau\| < \eta(\tau, x, \epsilon_1)$,

$$
(h_\tau(x) - h_{\tau'}(x'))^2 > \delta_1^2(\epsilon_1),
$$

(3.17)

where the function $\delta_1$, defined by

$$
\delta_1(y) = \frac{1}{4} \min \left( \left| \pi A \pi T e^{-\mu x} (1 - e^{-\mu y}) \right|, \left| \pi A \pi T e^{-\mu x} (e^{\mu y} - 1) \right| \right),
$$

(3.18)

is positive and is an increasing function in $y$ when $y > 0$.

We are now ready to establish the claim that the inverse mapping is continuous. For any $\epsilon > 0$, and $\|\xi' - \xi\| > \epsilon$, we consider the following two cases.

(a) If $\tau'$ and $\tau$ are close enough, and

$$
\|\tau' - \tau\| < \min \left( \frac{\epsilon}{2}, \eta(\tau, d_{13}, \frac{\epsilon}{2}), \eta(\tau, d_{23}, \frac{\epsilon}{2}) \right),
$$

(3.19)

then at least one of the following is true,

$$
|d'_{13} - d_{13}| > \frac{\epsilon}{2} \quad \text{or} \quad |d'_{23} - d_{23}| > \frac{\epsilon}{2}.
$$

Without loss of generality, suppose that

$$
|d'_{13} - d_{13}| > \frac{\epsilon}{2}.
$$

It follows from (3.17) that

$$
|h_1(\xi') - h_1(\xi)| > \delta_1(\frac{\epsilon}{2}).
$$

(3.20)
Note that $h_1$ is a sum of the four cell probabilities in $\mathbf{f}$ for which the nucleotide configurations share nucleotides A and T at nodes 1 and 3, and it follows that

$$
\|\mathbf{f}(\xi') - \mathbf{f}(\xi)\| \geq \frac{1}{2} |h_1(\xi') - h_1(\xi)| > \frac{1}{2} \delta_1 \left(\frac{\epsilon}{2}\right).
$$

(3.21)

(b) If $\tau'$ and $\tau$ are not very close, and

$$
\|\tau' - \tau\| \geq \min \left(\frac{\epsilon}{2}, \eta(\tau, d_{13}, \frac{\epsilon}{2}), \eta(\tau, d_{23}, \frac{\epsilon}{2})\right),
$$

then conditions for inequality (3.13) are met and it follows that

$$
\|\mathbf{f}(\xi') - \mathbf{f}(\xi)\| \geq \frac{1}{2} \|\mathbf{g}(\tau') - \mathbf{g}(\tau)\| \geq \frac{1}{2} \delta_2 \left(\min \left(\frac{\epsilon}{2}, \eta(\tau, d_{13}, \frac{\epsilon}{2}), \eta(\tau, d_{23}, \frac{\epsilon}{2})\right)\right). 
$$

(3.22)

2. Suppose the lemma holds true for any unrooted tree with $n$ external nodes.

Now consider an unrooted tree $T_1$ with $n + 1$ external nodes. (See Figure 3.1 for illustration.) Let $v_0$ be an external node (any external node), and $T_2$ denote the tree with $v_0$ removed from $T_1$. $\mathbf{B}$ denotes the branch lengths vector for $T_2$ with $t_{v_1v_2}$ removed. It follows from Proposition 3.1 that there exist two external nodes $b_1$ and $b_2$, and the branch lengths vector for $T_1$ in terms of edge lengths, $\mathbf{B} \| (t_{v_1v_3}, t_{v_2v_3}, t_{v_0v_3})$, differs from another parameterization $\mathbf{B} \| (t_{v_1v_2}, d_{v_0b_1}, d_{v_0b_2})$ by a linear transformation $M$ that is of full rank, and the metrics for the two parameterizations are equivalent.

Substituting $\tau\|\mathbf{B}$ into the role of $\tau$ in the proof for the case of a three-external-node tree, the similar proof can apply to tree $T_1$ with $n + 1$ external nodes. The only difference lies in that the probability of observing nucleotides A and T at nodes $v_0$ and $b_1$ is a sum of $4^{n-1}$ cell probabilities among $\mathbf{f}$, and the coefficients before $\delta_2$ in inequality (3.13) and $\delta_1$ in (3.21) need to be adjusted accordingly.
Thus, we have proved the continuity of $f^{-1}$ as claimed. 

So far we have verified all of Birch’s conditions, and thus proved the following theorem.

**Theorem 3.1.1** Consider an unrooted bifurcating tree with $N$ external nodes, whose tree topology is fixed and known. Let data $X$ represent the nucleotide sequences at the external nodes, each of which is $L$ nucleotides long. Assume also that the nucleotide evolution follows the F84 model with evolutionary parameters $\mu, K, \pi_A, \pi_G$, and $\pi_T$. Let $\mathbf{t}$ denote the branch lengths vector under the topology $T$. Denote the parameter in the model composed of evolutionary parameters and the branch lengths vector by $\theta = \tau || \mathbf{t} = (\mu, K, \pi_A, \pi_G, \pi_T) || \mathbf{t}$. Then as $L$ approaches infinity, the asymptotic distribution for $\sqrt{L}(\hat{\theta} - \theta)$, where $\hat{\theta}$ is the MLE of $\theta$, is multivariate normal with mean of vector $\mathbf{0}$ and variance-covariance matrix given by the inverse Fisher information, i.e.,

$$\sqrt{L}(\hat{\theta} - \theta) \rightarrow \mathcal{N}(0, (A'A)^{-1}) \text{ in distribution},$$

where $A'A$ is the Fisher information matrix, and the $(i, j)$ element is

$$(A'A)_{i, j} = \sum_{k=1}^{4N} \left( \frac{\partial \log(f_k(\theta))}{\partial \theta_i} \right) \left( \frac{\partial \log(f_k(\theta))}{\partial \theta_j} \right) f_k(\theta),$$

where $f_k(\theta)$ is the multinomial probability of the $k$-th site pattern under the true parameter $\theta$.

**Proof:** Birch’s Theorem 2.2.1 [9].

3.2 Limiting Distribution of Joint MLE when the Tree Topology is Estimated

In the last section, we have shown that with the tree topology fixed, the joint MLE of evolutionary parameter and branch lengths vector is efficient. Now we will
state the more general result when the tree topology is unknown and is estimated simultaneously. Since the tree topology is a discrete parameter from a finite space, consistency implies that the probability of recovering the true topology tends to 1 as the number of sites approaches infinity.

**Theorem 3.2.1** Let $T$ be the topology of an unrooted bifurcating tree with $N$ external nodes. The branch lengths vector $\mathbf{t}$ is congruent with the topology $T$, and $\tau = (\mu, K, \pi_A, \pi_G, \pi_T)$ denotes the evolutionary parameter. Let $\theta$ denote $(\mathbf{t}, \tau)$. Let data $X$ represent the nucleotide sequences at the external nodes, each of which is $L$ nucleotides long. Assume also that the nucleotide evolution follows the F84 model with evolutionary parameters $\tau$ under a tree of $T$ and $\mathbf{t}$. Let $(\hat{T}_L, \hat{\mathbf{t}}_L, \hat{\tau}_L)$ represent the joint MLE from data $X$ and denote $\hat{\theta}_L = (\hat{\mathbf{t}}_L, \hat{\tau}_L)$. As $L$ approaches $\infty$,

1. $\hat{T}_L \xrightarrow{P} T$, 

2. $\hat{\theta}_L$ is asymptotically efficient.

$$\sqrt{L}(\hat{\theta}_L - \theta) \rightarrow \mathcal{N}(0, (A'A)^{-1}) \text{ in distribution,}$$

where $A'A$ is the Fisher information matrix, and the $(i, j)$ element is

$$(A'A)_{i,j} = \sum_{k=1}^{4^N} \left( \frac{\partial \log(f_k(\theta))}{\partial \theta_i} \right) \left( \frac{\partial \log(f_k(\theta))}{\partial \theta_j} \right) f_k(\theta),$$

where $f_k(\theta)$ is the multinomial probability of the $k$-th site pattern under the true topology $T$ and parameter of branch lengths and evolutionary parameter $\theta$.

**Proof:**

1. This is Chang’s result [17], as described in Section 2.1.2. The MLE of the tree topology is consistent, and thus

$$\text{Prob}(\hat{T}_L = T) \rightarrow 1.$$ 

(3.23)
2. The convergence in probability is equivalent to the convergence almost surely for the discrete-valued tree topology parameter from a finite space.

If a tree topology estimate $\hat{T}_L$ is the same as the true topology $T$, $\hat{\theta}_L - \theta$ is well-defined. For any measurable set $A$ in $\mathbb{R}^{2N+1}$ (the dimension for the identifiable parameter of $\theta$ is $2N+1$), as $L \to \infty$, we have

$$\text{Prob}(\hat{T}_L = T, \sqrt{L}(\hat{\theta}_L - \theta) \in A)$$

$$= \text{Prob}(\sqrt{L}(\hat{\theta}_L - \theta) \in A | \hat{T}_L = T) \cdot \text{Prob}(\hat{T}_L = T)$$

$$\to \text{Prob}(\sqrt{L}(\hat{\theta}_L - \theta) \in A | \hat{T}_L = T), \quad (3.24)$$

by (3.23).

Note that

$$\text{Prob}(\sqrt{L}(\hat{\theta}_L - \theta) \in A)$$

$$\leq \text{Prob}(\sqrt{L}(\hat{\theta}_L - \theta) \in A, \hat{T}_L = T) + \text{Prob}(\hat{T}_L \neq T)$$

$$\to \text{Prob}(\sqrt{L}(\hat{\theta}_L - \theta) \in A, \hat{T}_L = T), \quad (3.25)$$

and

$$\text{Prob}(\sqrt{L}(\hat{\theta}_L - \theta) \in A) \geq \text{Prob}(\sqrt{L}(\hat{\theta}_L - \theta) \in A, \hat{T}_L = T). \quad (3.26)$$

It follows that

$$\text{Prob}(\sqrt{L}(\hat{\theta}_L - \theta) \in A)$$

$$\to \text{Prob}(\sqrt{L}(\hat{\theta}_L - \theta) \in A, \hat{T}_L = T)$$

$$\to \text{Prob}(\sqrt{L}(\hat{\theta}_L - \theta) \in A | \hat{T}_L = T). \quad (3.27)$$

Therefore, with the efficiency results shown in Theorem 3.1.1, we have proved Theorem 3.2.1.
The above proof rests on the fact that as the number of sites gets large, the true tree topology in the finite discrete space is almost surely found by the MLE. Thus the branch lengths vector is congruent with the true tree topology, and therefore we can examine the large-sample distributional property of the branch lengths vector. However, it is important to note that unless the true tree topology can be determined, it is pointless to discuss the variability associated with the estimated branch lengths vector, as the set of possible values of branch lengths vector is tied to the tree topology.

The above Theorem provides a nice guide for the distributional property of the joint MLE under a large sample. However, the inverse Fisher information cannot be readily used for approximating the variance of the evolutionary parameter. The estimated topology from a real data set will not necessarily be the true topology, and we have no knowledge to evaluate that either. In fact, in Simulation 3 that was described in the last Chapter, when nucleotide data at 231 sites were generated at 14 sequences following an underlying topology and the joint MLE was searched, the percentage of times when the true tree topology was found to be ML tree topology was low. In the parameter setting with \( \mu = 0.5 \) and \( K = 0.5 \), only about 33% of times was the true tree topology inferred as the ML tree, while data sets under other parameter settings yielded even lower percentages. It is expected that this percentage will decrease with an increasing number of sequences and thus an increasing number of possible phylogenies. The Fisher information matrix cannot be defined if the tree topology is not the true one. Other approaches need to be sought to estimate the variability of maximum likelihood estimator of the parameters.

In the next Chapter, bootstrap methods are proposed for estimating the variability of the MLE of evolutionary parameters.
CHAPTER 4

BOOTSTRAP ESTIMATORS OF VARIANCE FOR THE MODEL PARAMETER MLE

In the last two Chapters, we have shown the large-sample distributional property of the MLE of model parameters, both in the case with a known tree and in the case of an unknown tree, in which tree topology and the branch lengths vector are part of the parameter to be estimated. Simulation 3 in Chapter 2 demonstrates that the estimate of variance for the MLE of evolutionary parameters without taking into account the uncertainty associated with tree estimation is a severe underestimate and could not be used as a practical solution. Furthermore, although the asymptotic efficiency result for evolutionary parameters and the branch lengths vector is established in the last chapter, the computation of the asymptotic variance-covariance matrix requires the knowledge of the true tree topology. We need to seek a practical algorithm to estimate the variability of the substitution model parameter MLE to deal with the real data situation where the true tree topology is always unknown with the very rare exception of laboratory-grown sequences.

We will appeal to the bootstrap approach in this chapter for an estimator of the s.e. for the MLE of evolutionary parameters, when the tree needs to be simultaneously estimated. Bootstrap has been widely used in many areas of statistics to estimate the
variability of a statistic when its distribution is either impossible or impractical to tract [30, 33, 24]. Hall's book [51] focused on the theoretical basis for the bootstrap methodology, employing an Edgeworth expansion to establish the convergence results for bootstrap estimators. Among the first to apply bootstrap to the phylogenetic tree problem was Felsenstein [37], in which he assessed the confidence of an estimated clade of the tree. The question that biologists are interested in here concerns whether the specified OTUs and only these OTUs among the sequences to be studied share a common direct ancestor, or the presence or absence of a certain monophyletic group, called a clade. Felsenstein introduced the use of the nonparametric bootstrap to assess how often a perturbation of the data will result in an inferred tree topology that shares the specified clade as that from the original data set. Hillis and Bull [58] used simulation studies and a laboratory-generated phylogeny to assess bootstrap values of clade support as measures of repeatability and accuracy and considered precision of these estimates. Their criticisms were that the distribution of the bootstrap frequency was far wider than that of the true frequency and also the bootstrap estimates of repeatability were biased. Efron et al. [31] pointed out the incorrect interpretation of bootstrap proportions in [58] in which the roles of bootstrap samples and the truth were mixed up, and revealed that the bootstrap proportions were approximately the posterior probabilities under a flat prior. Furthermore, they presented a two-level bootstrap procedure, albeit much more computationally intensive, to obtain the bootstrap confidence level in the traditional hypothesis testing setting. Newton [88] proved the theoretical results of large dispersion for the bootstrap distribution using relative entropy, and provided an explanation for the observed bias of the
bootstrap estimators for the proportion of a clade in the phylogeny. But the developed theory could not be used for practical computation or for post hoc adjustment. Holmes [60] provided a survey of the use of bootstrap methods in the phylogenetic setting and hurdles it faces when compared with classical bootstrap methods. Soltis and Soltis [108] presented an overview of the state of application of bootstrap in phylogeny reconstruction from a systematic biologist and practitioner’s perspective. The vast majority of the applications of bootstrap in the phylogenetic context has been on parsimony and distance-based tree building methods, and of interest was the confidence of an internal clade in the tree estimated from a data set.

Bootstrap estimates are appropriate for many common statistics, including smooth functions of solutions to smooth estimating equations and thus most ML estimators (Davison and Hinkley [24]). This motivates an approach that obtains the estimate of the variability of MLEs of evolutionary parameters when the tree is unknown and needs to be simultaneously estimated according to the traditional bootstrap scheme.

In this Chapter, we will first estimate the nucleotide equilibrium frequency parameters at the beginning of the search with the original data set, and then fix them for the estimation of the tree and other substitution model parameters, so the substitution model is in effect a two-parameter model with nucleotide frequency parameters fixed in the F84 model. This is due to the limitation in all of phylogenetic tree building software packages that are being used in practice as described in previous chapters. It is important to note that the bootstrap methods in this chapter can be readily adapted and extended to the nucleotide frequency parameters if a tree building program finds the joint MLE for equilibrium nucleotide frequency parameters along with other parameters.
4.1 Direct Bootstrap

Similar to the ordinary bootstrap scheme in phylogenetic data analysis, a direct bootstrap (DB) idea is described as follows.

Suppose the data set of interest has $N$ sequences, each of which is $L$ nucleotides long. Consider the matrix representation of the data set so that the data is displayed as an $N \times L$ matrix $x$. We draw columns of the data matrix $x$ with replacement $L$ times to form a new data set $x_i^*$. For $x_i^*$, we simultaneously estimate the tree topology, branch lengths and the evolutionary parameters, and denote the MLE for $\mu$ and $K$ as $\mu_i^*$ and $K_i^*$. We repeat the above process $B$ times, and the empirical variances based on the sets of $\mu_i^*, i = 1, \ldots, B$ and $K_i^*, i = 1, \ldots, B$ provide the bootstrap estimates of the variances for the MLE.

We note that this procedure corresponds to a nonparametric bootstrapping approach, as it involves simply resampling from the columns of the original data matrix. Similarly a parametric bootstrap can also be proposed. Instead of generating the bootstrap sample by resampling the columns in the data matrix, the parametric bootstrapping scheme draws $L$ i.i.d. $N$-dimensional nucleotide vectors from a parametric distribution determined from the data $x$ and forms a bootstrap sample $x_i^*$. In particular, the parametric distribution of the nucleotide vector is the multinomial distribution that is determined by the MLE of tree topology, branch lengths and evolutionary parameters combined. In the phylogenetic tree setting, we will draw $B$ bootstrap samples from the distribution according to the estimated tree and evolutionary parameters. For each bootstrap sample $x_i^*, i = 1, \ldots, B$, we will generate $L$ sites of nucleotide data at the $N$ external nodes of the tree according to the substitution model with estimated evolutionary parameters and the estimated tree. The
bootstrap variance estimate for the MLE of evolutionary parameters can be obtained following the same procedure as described for the nonparametric bootstrap.

We note the similarity of the process of drawing samples in the parametric bootstrap and the MCS method. Nucleotide data are generated in the same manner at a single site. However, for the MCS method, the number of sites to be generated needs to be large to approximate the probability distribution. The larger the number of sites of data it generates, the closer the empirical c.d.f. is to the true multinomial distribution and thus the better. The parametric bootstrap will draw nucleotide data to form only the same number of columns as there is in the original data set to approximate the variability associated with the MLE from such a data set.

Both direct bootstrap methods require that, for each of the $B$ bootstrap samples, the tree topology, branch lengths and the evolutionary parameters be simultaneously estimated by optimizing the likelihood function, and are thus computationally expensive. The computational cost for finding the ML tree given the evolutionary parameters and for finding the MLE of evolutionary parameters given the tree is substantially lower, which motivates an alternative bootstrap idea.

4.2 Conditional Variance Bootstrap

To illustrate the idea, we consider the F84 model with the equilibrium nucleotide frequency parameters fixed as was commonly done in the implementation of searches for the MLE. The proposed bootstrap method that follows can be applied to the case where the MLE of equilibrium nucleotide frequency parameters are simultaneously searched.
Let $\hat{\theta}$ denote the maximum likelihood estimates of the evolutionary parameter $\mu$ and $K$, and $t_\theta(x)$ denote the maximum likelihood tree topology and vector of branch lengths as the evolutionary parameter is fixed at $\hat{\theta}$. For the sake of argument, we take the parameter $\mu$ as an example in the following derivation, and the formulations for $K$ are the same. Let $f_{t_\theta(x)}(x)$ denote the maximum likelihood estimate of $\mu$ with the tree fixed at $t_\theta(x)$. By the conditioning principle,

$$Var(\hat{\mu}) = E(Var(f_{t_\theta(x)}(x)|t_\theta(x))) + \text{Var}(E(f_{t_\theta(x)}(x)|t_\theta(x))). \quad (4.1)$$

Following the standard bootstrap derivation, we have

$$\text{Var}_F(\hat{\mu}) = E_F(Var(f_{t_\theta(x^*)}(x^*)|t_\theta(x^*))) + \text{Var}_F(E(f_{t_\theta(x^*)}(x^*)|t_\theta(x^*))), \quad (4.2)$$

where $F$ denotes the empirical c.d.f. of the multinomial distribution from data $x$ following the approach of the nonparametric bootstrap, and $x^*$ denotes a bootstrap sample formed by drawing the columns of $x$ with replacement $L$ times. The expression $t_\theta(x^*)$ represents the ML tree estimated from the data of the bootstrap sample $x^*$ while the evolutionary parameters are fixed at $\hat{\theta}$, the global MLE of $\theta$, for the original data $x$.

To motivate a practical algorithm, we note that the MCS method can provide an approximation to the asymptotic variance of ML estimates of evolution parameters when the tree is fixed. Thus, the MCS estimates to the conditional variance inside the expectation of the first term on the right-hand side (RHS) of (4.2) can be obtained and then averaged over all $B$ bootstrap samples to provide the expectation term. The MLE $f_{t_\theta(x^*)}(x^*)$ itself is employed so as to approximate the conditional expectation of the second term on the right-hand side. The details of the algorithm are as follows:
1. From the data set $x$, obtain the global maximum likelihood estimate of evolutionary parameters, $\hat{\theta}$, and that of the tree $t_\theta(x)$.

2. Generate $B$ non-parametric bootstrap samples, each by sampling the columns of $x$ with replacement $L$ times.

3. For each of the bootstrap sample $x_i^*$ obtained in Step 2, find the ML tree $t_\theta(x_i^*)$ with the evolutionary parameters fixed at the value $\hat{\theta}$ determined in Step 1.

4. With the tree fixed at this tree obtained in Step 3, find the ML estimates of evolutionary parameters, $\hat{\mu}_i$ and $\hat{K}_i$. These serve as estimates to approximate the conditional expectations in the second term of the RHS of Equation (4.2).

5. For each tree $t_\theta(x_i^*)$, $i = 1, \ldots, B$, with parameters fixed at $\hat{\theta}$, generate $nL$ sites ($n$ an integer). The conditional variance in the first term of the RHS of Equation (4.2) is estimated by the inverse of the expected Fisher information following the MCS method as described in Chapter 2. Denote these estimates of conditional variances by $\tilde{v}_i^*(\mu)$ and $\tilde{v}_i^*(K)$.

6. The averages of $\tilde{v}_i^*(\mu)$ and $\tilde{v}_i^*(K)$ are computed to get the first term in Equation (4.2), and the sample variances of the $\hat{\mu}_i^*$ and the $\hat{K}_i^*$ are computed to get the second term in Equation (4.2). The estimate of the variance for $\hat{\mu}$ is then the sum of these two quantities for $\mu$, and the estimate of the variance for $\hat{K}$ is the sum of these two respective quantities for $K$.

The above algorithm takes advantage of the time savings with estimating the tree when evolutionary parameters are fixed and with estimating evolutionary parameters when the tree is fixed.
In Step 2 of the algorithm, if the bootstrap samples are instead generated from the multinomial distribution determined by the global ML tree $t_{\hat{\theta}}(x)$ and MLE for evolutionary parameters, $\hat{\theta}$, we have a parametric bootstrap algorithm.

The replacement of the expectation of the evolutionary parameter MLE by the value of the MLE, as was done in Step 4, is a crude estimate. However, it contains arguably all of information of the conditional distribution. Efron and Tibshirani [33] have introduced the idea of bias-corrected bootstrap methods. In the ML phylogenetic setting, a bias-corrected estimate of the conditional expectation based on the bootstrap can be obtained as follows. For the $nL$ sites generated in Step 5, separate them into $n$ groups, each consisting of data with $L$ sites. For each of the $n$ data sets, with the tree fixed at the same tree that was used to generate the nucleotide data at these $nL$ sites, find the MLEs of $\mu$ and $K$. Denote the average of the $n$ estimates by $\bar{v}_i^*(\mu)$ and $\bar{v}_i^*(K)$ for $\mu$ and $K$, respectively. Then the bias-corrected version to the conditional expectation in the second term of RHS of (4.2) is $2\bar{v}_i^*(\mu) - \bar{v}_i^*(\mu)$ and $2\bar{v}_i^*(K) - \bar{v}_i^*(K)$. The bias correction, however, is known to result in a larger variance associated with the estimates in general [33].

4.3 Simulation

We next examine and evaluate the performances of the direct bootstrap (DB), the bootstrap method motivated by the conditional variance principle (CVB), and the bias-corrected version (BC-CVB). Once again, we evaluate the variance estimators for the MLE of evolutionary parameters via simulations, with four parameter settings of $(\mu, K)$ and true tree as the ML tree from the mtDNA real data set.
In addition to the above three bootstrap methods, the performance of another estimator of the variability is evaluated as well. In Wang et al. [125], the method to approximate the conditional expectation in the second term on the right-hand side of Equation (4.2) is to generate multiple simulated data sets with $L$ sites and take the average of the resulting MLE of $\mu$ and $K$ from these simulated data. The recommended option in the paper is to generate simulated data at the parameter estimate from the bootstrap sample, and it is named CVB(BP).

For each setting of $(\mu, K) = (0.1, 0.5), (0.1, 10), (0.5, 0.5)$ and $(0.5, 10)$, with the tree fixed at the ML tree from mtDNA data (Figure 2.2), we generate 10 sets of nucleotide data, each of which consists of 14 sequences and 231 sites in each sequence. For each set, we follow the three bootstrap methods of DB, CVB and BC-CVB described in the last two sections to obtain the variance estimates, as well as the CVB(BP) method. For each bootstrap method, both versions of parametric and nonparametric approach were carried out. Fifty (50) bootstrap replicates were used for each simulation run of variance estimation. Once again, Seq-Gen [95] was used to draw nucleotides for parametric bootstrap samples and samples for the MCS method. Results of the nonparametric bootstrap are summarized in Tables 4.1 and 4.2, and results of parametric bootstrap are summarized in Tables 4.3 and 4.4.
Table 4.1: Mean and standard deviation of the simulation results for MLE of $\mu$ using nonparametric DB, CVB, BC-CVB and CVB(BP) methods

<table>
<thead>
<tr>
<th>$\mu$</th>
<th>$K$</th>
<th>Simulation 3 SD($\mu$)</th>
<th>DB</th>
<th>CVB</th>
<th>BC-CVB</th>
<th>CVB(BP)</th>
</tr>
</thead>
<tbody>
<tr>
<td>0.1</td>
<td>0.5</td>
<td>0.0145</td>
<td>Mean 0.0153</td>
<td>0.0193</td>
<td>0.0194</td>
<td>0.0198</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>(s.d.) (0.0021)</td>
<td>(0.0022)</td>
<td>(0.0022)</td>
<td>(0.0023)</td>
</tr>
<tr>
<td>0.1</td>
<td>10.0</td>
<td>0.0148</td>
<td>Mean 0.0135</td>
<td>0.0177</td>
<td>0.0184</td>
<td>0.0182</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>(s.d.) (0.0030)</td>
<td>(0.0031)</td>
<td>(0.0033)</td>
<td>(0.0033)</td>
</tr>
<tr>
<td>0.5</td>
<td>0.5</td>
<td>0.0492</td>
<td>Mean 0.0454</td>
<td>0.0516</td>
<td>0.0529</td>
<td>0.0520</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>(s.d.) (0.0114)</td>
<td>(0.0099)</td>
<td>(0.0098)</td>
<td>(0.0102)</td>
</tr>
<tr>
<td>0.5</td>
<td>10.0</td>
<td>0.0572</td>
<td>Mean 0.0621</td>
<td>0.0665</td>
<td>0.0778</td>
<td>0.0640</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>(s.d.) (0.0109)</td>
<td>(0.0096)</td>
<td>(0.0208)</td>
<td>(0.0076)</td>
</tr>
</tbody>
</table>

Table 4.2: Mean and standard deviation of the simulation results for MLE of $K$ using nonparametric DB, CVB, BC-CVB and CVB(BP) methods

<table>
<thead>
<tr>
<th>$\mu$</th>
<th>$K$</th>
<th>Simulation 3 SD($\mu$)</th>
<th>DB</th>
<th>CVB</th>
<th>BC-CVB</th>
<th>CVB(BP)</th>
</tr>
</thead>
<tbody>
<tr>
<td>0.1</td>
<td>0.5</td>
<td>0.1702</td>
<td>Mean 0.1638</td>
<td>0.2382</td>
<td>0.2404</td>
<td>0.2478</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>(s.d.) (0.0429)</td>
<td>(0.0501)</td>
<td>(0.0508)</td>
<td>(0.0511)</td>
</tr>
<tr>
<td>0.1</td>
<td>10.0</td>
<td>1.4135</td>
<td>Mean 1.5019</td>
<td>2.1266</td>
<td>2.3867</td>
<td>2.1335</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>(s.d.) (0.3150)</td>
<td>(0.3803)</td>
<td>(0.5211)</td>
<td>(0.4029)</td>
</tr>
<tr>
<td>0.5</td>
<td>0.5</td>
<td>0.0885</td>
<td>Mean 0.1031</td>
<td>0.1478</td>
<td>0.1522</td>
<td>0.1514</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>(s.d.) (0.0487)</td>
<td>(0.0699)</td>
<td>(0.0728)</td>
<td>(0.0697)</td>
</tr>
<tr>
<td>0.5</td>
<td>10.0</td>
<td>2.5067</td>
<td>Mean 2.9305</td>
<td>3.4137</td>
<td>3.4824</td>
<td>3.5386</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>(s.d.) (1.5593)</td>
<td>(1.6300)</td>
<td>(1.7123)</td>
<td>(1.5741)</td>
</tr>
</tbody>
</table>

Table 4.3: Mean and standard deviation of the simulation results for MLE of $\mu$ using parametric DB, CVB, BC-CVB and CVB(BP) methods

<table>
<thead>
<tr>
<th>$\mu$</th>
<th>$K$</th>
<th>Simulation 3 SD($\mu$)</th>
<th>DB</th>
<th>CVB</th>
<th>BC-CVB</th>
<th>CVB(BP)</th>
</tr>
</thead>
<tbody>
<tr>
<td>0.1</td>
<td>0.5</td>
<td>0.0145</td>
<td>Mean 0.0146</td>
<td>0.0187</td>
<td>0.0191</td>
<td>0.0192</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>(s.d.) (0.0037)</td>
<td>(0.0036)</td>
<td>(0.0035)</td>
<td>(0.0037)</td>
</tr>
<tr>
<td>0.1</td>
<td>10.0</td>
<td>0.0148</td>
<td>Mean 0.0148</td>
<td>0.0191</td>
<td>0.0196</td>
<td>0.0194</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>(s.d.) (0.0030)</td>
<td>(0.0028)</td>
<td>(0.0033)</td>
<td>(0.0026)</td>
</tr>
<tr>
<td>$\mu$</td>
<td>$K$</td>
<td>SD($\mu$)</td>
<td>Simulation 3</td>
<td>DB</td>
<td>CVB</td>
<td>BC-CVB</td>
</tr>
<tr>
<td>------</td>
<td>-----</td>
<td>-----------</td>
<td>-------------</td>
<td>----</td>
<td>-----</td>
<td>--------</td>
</tr>
<tr>
<td>0.5</td>
<td>0.5</td>
<td>0.0492</td>
<td>Mean</td>
<td>0.0421</td>
<td>0.0495</td>
<td>0.0506</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>(s.d.)</td>
<td>(0.0120)</td>
<td>(0.0092)</td>
<td>(0.0092)</td>
</tr>
<tr>
<td>0.5</td>
<td>10.0</td>
<td>0.0572</td>
<td>Mean</td>
<td>0.0530</td>
<td>0.0602</td>
<td>0.0683</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>(s.d.)</td>
<td>(0.0129)</td>
<td>(0.0108)</td>
<td>(0.0221)</td>
</tr>
</tbody>
</table>

Table 4.3: Mean and standard deviation of the simulation results for s.e. of MLE of $\mu$ using parametric DB, CVB, BC-CVB and CVB(BP) methods

<table>
<thead>
<tr>
<th>$\mu$</th>
<th>$K$</th>
<th>SD($\mu$)</th>
<th>Simulation 3</th>
<th>DB</th>
<th>CVB</th>
<th>BC-CVB</th>
<th>CVB(BP)</th>
</tr>
</thead>
<tbody>
<tr>
<td>0.1</td>
<td>0.5</td>
<td>0.1702</td>
<td>Mean</td>
<td>0.1644</td>
<td>0.2417</td>
<td>0.2467</td>
<td>0.2495</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>(s.d.)</td>
<td>(0.0242)</td>
<td>(0.0372)</td>
<td>(0.0393)</td>
<td>(0.0364)</td>
</tr>
<tr>
<td>0.1</td>
<td>10.0</td>
<td>1.4135</td>
<td>Mean</td>
<td>1.2937</td>
<td>1.8732</td>
<td>2.0931</td>
<td>1.9273</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>(s.d.)</td>
<td>(0.3344)</td>
<td>(0.3934)</td>
<td>(0.6124)</td>
<td>(0.4162)</td>
</tr>
<tr>
<td>0.5</td>
<td>0.5</td>
<td>0.0885</td>
<td>Mean</td>
<td>0.0900</td>
<td>0.1271</td>
<td>0.1296</td>
<td>0.1309</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>(s.d.)</td>
<td>(0.0097)</td>
<td>(0.0114)</td>
<td>(0.0111)</td>
<td>(0.0134)</td>
</tr>
<tr>
<td>0.5</td>
<td>10.0</td>
<td>2.5067</td>
<td>Mean</td>
<td>2.2132</td>
<td>2.8482</td>
<td>2.9555</td>
<td>2.9827</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>(s.d.)</td>
<td>(1.0082)</td>
<td>(1.2542)</td>
<td>(1.4089)</td>
<td>(1.2708)</td>
</tr>
</tbody>
</table>

Table 4.4: Mean and standard deviation of the simulation results for s.e. of MLE of $K$ using parametric DB, CVB, BC-CVB and CVB(BP) methods

The results from the simulations can be further summarized by a measure similar to the ordinary MSE. Pretending that the standard deviations of MLEs of $\mu$ and $K$ from Simulation 3 were the true standard deviations, we calculate the mean squared error of the s.e. estimates by the eight bootstrap methods for each of the ten simulated data sets at each of the four parameter settings.
<table>
<thead>
<tr>
<th>( \mu )</th>
<th>K</th>
<th>Nonparametric</th>
<th>Parametric</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td>DB</td>
<td>CVB</td>
</tr>
<tr>
<td>0.1</td>
<td>0.5</td>
<td>0.484(^1)</td>
<td>2.29</td>
</tr>
<tr>
<td>0.1</td>
<td>10.0</td>
<td>1.07(^2)</td>
<td>1.77</td>
</tr>
<tr>
<td>0.5</td>
<td>0.5</td>
<td>14.4</td>
<td>10.4</td>
</tr>
<tr>
<td>0.5</td>
<td>10.0</td>
<td>14.4</td>
<td>17.9</td>
</tr>
</tbody>
</table>

Table 4.5: Summary of pseudo-MSE of bootstrap estimate for \( \mu \) of s.e. by the eight methods, pretending that s.d. from Simulation 3 was the true value. Entries need to be multiplied by \(10^{-6}\) for pseudo-MSEs. Entries marked with superscripts 1 and 2 have the smallest and second smallest pseudo-MSEs.

<table>
<thead>
<tr>
<th>( \mu )</th>
<th>K</th>
<th>Nonparametric</th>
<th>Parametric</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td>DB</td>
<td>CVB</td>
</tr>
<tr>
<td>0.1</td>
<td>0.5</td>
<td>0.0188(^1)</td>
<td>0.0713</td>
</tr>
<tr>
<td>0.1</td>
<td>10.0</td>
<td>1.07(^1)</td>
<td>6.53</td>
</tr>
<tr>
<td>0.5</td>
<td>0.5</td>
<td>0.0259</td>
<td>0.0840</td>
</tr>
<tr>
<td>0.5</td>
<td>10.0</td>
<td>26.1</td>
<td>34.8</td>
</tr>
</tbody>
</table>

Table 4.6: Summary of pseudo-MSE of bootstrap estimate for \( K \) of s.e. by the eight methods, pretending that s.d. from Simulation 3 was the true value. Entries need to be multiplied by \(10^{-2}\) for pseudo-MSEs. Entries marked with superscripts 1 and 2 have the smallest and second smallest pseudo-MSEs.

Upon examining the results in Tables 4.1 through 4.6, we find that both the direct bootstrap and conditional variance bootstrap methods give reasonably good estimators of the variability of MLEs for evolutionary parameters. The BC-CVB bootstrap method, supposedly correcting for the bias of the conditional expectation term, is expected to produce a larger variance, and thus enlarge the second term of the right-hand side of Equation (4.2). In turn, the bootstrap estimate of the variability by this method is an overestimate as expected, and the CVB method outperforms it in each and every setting. CVB(BP) method generates simulated data and employs...
the average of the estimates to approximate the conditional expectation of the MLE of model parameters given the tree. However, the approach does not have sound justifications and the simulation performance shows that it is associated with an upward bias and is in general inferior to the CVB method with the exception of estimating \( \mu \) at the setting of \( (\mu, K) = (0.5, 10) \).

Neither the parametric nor nonparametric DB method shows discernible pattern as to under-estimation or over-estimation of the variability of the MLE of evolutionary parameters, when compared to the benchmark of variability observed in Simulation 3. According to the pseudo-MSE criterion, the parametric DB method performs slightly better in general, as is the case for CVB method.

Both of the CVB estimators appear to be larger in general than the value from Simulation 3. The parametric CVB performs rather well for the cases of \( \mu = 0.5 \), the high setting. CVB method seems to give a slightly biased estimator of the variability and it tends to overestimate.

Another factor that demands much consideration in practice is the run time cost, as the tree searching with the ML method is time consuming. For the nonparametric bootstrap, the CVB method demonstrated run time savings ranging from about 15% to 30%; while in the recommended parametric bootstrap case, the saving ranged from 20% to 38%, with a mean of about 30%. The time saving is expected to increase with the larger number of OTUs and the complexity of the data set.

Summarizing these simulations, the two bootstrap methods, direct bootstrap (DB) and conditional variance bootstrap (CVB) appear to be better. The parametric bootstrap in general outperforms the nonparametric counterpart. The CVB method performs reasonably well, and appears to give a more conservative estimator of the
variability. It performs well for the cases when the true overall substitution rate is higher. The time saving of CVB over DB is expected to be substantial with a large complex data set. In practice, if the run time is not a concern, (for example, a smaller data set), the DB method is recommended, while CVB is a good method for large data sets and when a more conservative estimate of the variability can be accepted.

In the next section, the CVB method is to be applied to a real data set of HPV sequences.

4.4 Application to HPV Sequences

Cervical cancer is the second most common cancer in women worldwide, and it is recognized from molecular epidemiologic evidence that certain types of human papillomavirus (HPV) are the principal cause of invasive cervical cancer and cervical intra-epithelial neoplasia. In various epidemiologic studies over a wide range of countries and populations, presence of HPV has been reported in 90% to 95% of cervical cancers, as compared to about 5% in asymptomatic women of similar age. Among the more than 100 types of papillomaviruses identified, approximately 40 infect the genital tract. Moreover, although nearly all cervical cancers contain HPV genomes, most HPVs are not oncogenic but are only associated with skin warts or benign tumors. Burk [12] provided an overview of the subject. It is important to identify the high-risk and low-risk groups of HPVs for screening programs and HPV vaccine development. Epidemiologic studies, most notably Muñoz et al. [85], have reported the high-risk types, but this type of studies is expensive, and may not be conclusive for the types with low prevalence. The phylogenetic classification is less expensive, and
provides useful information for evolutionary proximity and it can be used to compare
with the findings in epidemiologic studies.

Aligned DNA sequences for 30 papillomaviruses (28 human papillomaviruses (HPVs),
a rhesus papillomavirus (RhPV), and a pygmy chimpanzee papillomavirus (PCPV))
were obtained from the Los Alamos National Database website (http://hpv-web.lanl.gov).
After removal of sites with any insertion or deletion occurring in any sequence, the
resulting data set contains 1,382 base pairs for the L1 gene. This data set has been
studied by Chan et al. [15, 14] and Ong et al. [90]. In particular, Ong et al. [90]
concluded that the molecular clock assumption is appropriate for this data set. Here
we consider the problem of simultaneously estimating the tree and the parameters \( \mu \)
and \( K \) in the context of the F84 model [36]. As was done in the previous sections,
the empirical frequency of nucleotides is taken as the fixed equilibrium nucleotide
frequency parameter in the model, and this is purely because of the limitation of the
packages that implement the tree building procedures. The only two evolutionary
parameters remaining in the F84 model are then \( \mu \) and \( K \).

Common phylogenetic analysis packages, such as PHYLIP and PAUP*, use a dif-
fferent parameterization for the F84 model, and in particular, the transition/transversion
ratio parameter \( R \) is used which is the ratio of the instantaneous rates of the two.
\( R \) and \( K \) can be converted to each other, but the conversion formula depends on
the equilibrium nucleotide frequency parameters. Because the estimation of the ad-
ditional model parameters adds quite substantial computational costs, both PHYLIP
and PAUP* give an option of setting \( R \) to the default value of 2.0. For this fairly
large data set with 30 sequences, it is tempting to set \( R \) at the default value 2.0,
and that corresponds to the \( K \) value in the F84 model of 1.5691. Under this fixed
default value of $K$, SSA was used to determine the ML tree, labeled Tree $a$, as shown in Figure 4.1 (a). With the tree fixed at Tree $a$, an optimization was performed on $(\mu, K)$, and the values of 0.3010 and 0.6606 resulted, under which the log likelihood of Tree $a$ was -27,864.73. In practice, this one-step iterative procedure is commonly performed by estimating the tree at the default value of evolutionary model parameter(s) and then estimate the evolutionary parameters with the tree fixed at the estimated tree. Although the estimated $K$ value of 0.6606 and the initial default value of 1.5691 differ considerably, the variability associated with the estimation needs to be quantified. Under $(\mu, K) = (0.3010, 0.6606)$ and Tree $a$, twenty (20) parametric bootstrap samples were generated and the CVB method was carried out to derive an estimated s.e. for $\mu$ and $K$ using the procedure described in the last section. The MCS method was executed by generating nucleotide data at sites of ten times the length of the original data set. It resulted in an estimated $s.e.(\hat{K}) = 0.0428$, so an approximate 95% confidence interval for $K$ is $(0.5767, 0.7449)$, which shows that the original data set contain information to estimate $K$ rather well and it does not support the default value $K$ of 1.5691. It suggests, in this case, the simultaneous estimation of the transition/transversion parameter along with other parameters in the model may be important.

Next, tree topology, branch lengths and both evolutionary parameters $\mu$ and $K$ were simultaneously estimated by SSA. The resulting ML tree is shown in Figure 4.1 (b), and labeled as Tree $b$. The ML estimate of $(\mu, K)$ is $(0.3014, 0.6578)$, under which Tree $b$ has a log likelihood of -27,864.11. Again, twenty (20) parametric bootstrap samples were generated and the CVB method was carried out to estimate the s.e. of the evolutionary parameters. The estimated $s.e.(\hat{K}) = 0.0533$,
Figure 4.1: ML trees for papillomavirus data set. Tree a is the ML tree with the transition/transversion ratio $R$ fixed at 2.0, and the log likelihood with MLE of $(\mu, K) = (0.3010, 0.6606)$ under this tree is -27,864.73. Tree b is the ML tree when the transition/transversion ratio $R$ is simultaneously estimated, and the log likelihood with MLE of $(\mu, K) = (0.3014, 0.6578)$ under this tree is -27,864.11. Bold-face sequences represented known high-risk oncogenic strains. Note the different classifications for strain HPV34.
and an approximate 95% confidence interval is (0.5533, 0.7623), which suggests that the transition/transversion ratio in this example is rather small and less than 1. The estimated s.e. along with the MLE of $\mu$ and $K$ by the two approaches of fixing $K$ at the default value to begin with and simultaneously estimating $K$ are presented in Table 4.7.

<table>
<thead>
<tr>
<th>Tree estimated assuming:</th>
<th>$\hat{\mu}$</th>
<th>$s.e.(\hat{\mu})$</th>
<th>$\hat{K}$</th>
<th>$s.e.(\hat{K})$</th>
</tr>
</thead>
<tbody>
<tr>
<td>R fixed (2.0)</td>
<td>0.3010</td>
<td>0.0124</td>
<td>0.6606</td>
<td>0.0428</td>
</tr>
<tr>
<td>R estimated</td>
<td>0.3014</td>
<td>0.0132</td>
<td>0.6578</td>
<td>0.0533</td>
</tr>
</tbody>
</table>

Table 4.7: MLE of $\mu$ and $K$ and CVB estimates of their s.e.’s for the papillomavirus data set. Two ML trees were obtained with transition/transversion ratio $R$ fixed at the default value 2.0 and with $R$ simultaneously estimated, respectively. The ML estimates of $\mu$ and $K$ were obtained with respect to each of the two ML trees. Twenty (20) parametric bootstrap samples were generated under each of the two sets of tree and evolutionary parameters, and the estimated s.e.’s from CVB method are reported.

The oncogenic classification in Burk [12] shows that among the sequences considered in this example, sequences HPV16, HPV18, HPV45, HPV31, HPV26, HPV33, HPV35, HPV39, HPV51, HPV53, HPV56, HPV58, and HPV59 are classified as high-risk papillomaviruses associated with causing cervical cancer. (Only high-risk papillomaviruses were reported in [12], while the low-risk HPVs were not listed.) These sequences are highlighted in bold face in Figure 4.1. The molecular epidemiologic study by Muñoz et al. [85] has confirmed the above classification, and in addition, HPV40 was reported to be in low-risk group. Some of the sequences, e.g., HPV34 and HPV7, were not mentioned in the study, due possibly to the absence or low prevalence in the subjects of this observational study. If we compare Tree a to the “true” ML tree Tree b, the only difference lies in the placement of the sequence HPV34. In Tree
a, HPV34 is clustered with the clade of six known oncogenic papillomaviruses. In contrast, in Tree b, HPV34 is clustered first with HPV7 and HPV40, and the latter is known to be a low-risk papillomavirus. Thus, if the risk of oncogenicity of HPV34 is to be inferred based on the phylogenetic information alone, these two trees will lead to different answers with respect to HPV34. (Classification of pathogenicity of viruses was studied under a given phylogenetic tree in [124].)

In phylogenetic analysis, it is also of interest to investigate other trees with high likelihood close to that of the ML tree. Further examination reveals that there is another island of three trees with high likelihood when the transition/transversion ratio parameter $R$ is fixed at the default value 2.0. See Salter [101] for details. This island is often found by the heuristic uphill search algorithms, such as PHYLIP and PAUP*. The trees c, d, and e are shown in Figure 4.2. When $R$ is fixed at 2.0, they have log likelihood of -28,086.26 to -28,086.23, as compared to the log likelihood for Tree a of -28,083.67 and that for Tree b of -28,084.42. Upon a closer examination of the three trees, they differ from one another by an NNI (nearest neighbor exchange), where the NNI concept was shown in Figure 1.2 [69] in Chapter 1. In fact, these three trees share the same three clades, \{HPV57, HPV2a, HPV27\}, \{HPV32, HPV42, ..., RhPV1\}, and \{HPV3, HPV10, ..., HPV45\}, and they are the three possible trees from these clades. Note that HPV34 is clustered with RhPV1 in all three phylogenies, and then they are clustered with a clade consisting of oncogenic sequences. When the phylogeny of Tree c is compared with those of Tree a and Tree b, the only difference concerns the placement of the two sequences of HPV34 and RhPV1, but Tree c is separated from Tree a or Tree b by more than 5 NNIs.
Figure 4.2: Other high-likelihood trees for papillomavirus data set. These trees have high likelihood values when $R$ is fixed at the default value of 2.0. The three trees differ by only one NNI from each other, that is, they represent the three possible trees of the three clades. Sequences in bold face are the known oncogenic strains.
When the transition/transversion ratio parameter is simultaneously optimized, however, the likelihood between the island of Trees c, Tree d and Tree e and the other two Trees a and b differ by a much greater margin. The values of log likelihood of the three trees are all very near -27,871.83, about 6 or 7 units from that for Tree b or a, as opposed to about 3 units in the case of fixed K. This demonstrates that simultaneous estimation of the transition/transversion ratio parameter for this data set is important to obtain the “true” ML phylogeny, and also to investigate and understand phylogenies close to the global optimum. The estimate of variability of K allows us to demonstrate that data often does not support the common practice of setting it to the default value.

4.5 Discussion

In this thesis, we have shown the large-sample property for the joint maximum likelihood estimator of tree topology, branch lengths vector, and evolutionary parameters under perhaps the most commonly used substitution models, namely the F84 or HYK85 model. The probability of recovering the true tree topology goes to 1 as the number of sites approaches \( \infty \). The difference between the MLE and the true value for the evolutionary parameters and branch lengths vector times the square root of the number of sites approaches a multivariate normal distribution with the mean vector of 0 and variance-covariance matrix of the inverse of Fisher information evaluated at the true parameter value. In addition to the point estimate from the common phylogenetic analysis packages that find a set of tree topology, branch lengths vector and evolutionary parameter values to maximize the likelihood under a pre-specified
substitution model, it is important to gain some idea of the variability of the estimate. For the tree topology, bootstrapping has been widely implemented although the interpretation of the results needs caution [31]. For estimating the variability of the MLE for evolutionary parameters, the approach is described as follows.

In the situation where the tree topology and branch lengths vector are considered known, the estimate of variability of the evolutionary parameter can be derived from the inverse Fisher information following Theorem 2.2.2. Based on simulation, the estimator of expected Fisher information is demonstrated to outperform the observed Fisher information that was implemented in the only known package to give such a variability estimate. The MCS method to generate nucleotide data with the evolutionary parameters fixed at the MLE at a number of sites proves to be a good practical algorithm to approximate the expected Fisher information. The recommended number of sites for which nucleotide data to be generated is 10 to 20 times the length of the sequences in the original data set.

In the situation where the tree topology can be considered known, but branch lengths are unknown, similar to the above, an MCS algorithm can be carried out by generating simulated nucleotide data at a number of sites under the MLE of evolutionary parameters and branch lengths vector and the fixed topology. Although not being tested by simulation, it is expected that this approach would outperform that of observed Fisher information. The method can provide an estimate of variability associated with the estimated branch lengths as well.

In most of the real situations, tree topology is unknown and needs to be estimated. Through simulation, a severe underestimation of the variability for the evolutionary parameters is demonstrated when ignoring variability associated with the uncertainty.
stemming from the estimation of tree topology and branch lengths. The tree topology estimator, although large-sample consistency holds true, is rarely the true topology for a real data set, and the Fisher information cannot be directly calculated. Bootstrap methods are proposed for estimation of the variability of MLE. The standard direct bootstrap (DB) can be carried out by drawing sets of bootstrap samples, obtaining MLEs for these sets, and finding the empirical variability of these MLEs across various sets. DB method performs well, however, the computational cost that is involved by simultaneously estimating tree and evolutionary parameters is high and this approach may not be practical for a large data set. The CVB method, motivated by the conditional variance principle, has been shown by simulation to perform reasonably well. It is observed to produce a slight upward bias for the variance estimate, which would correspond to a more conservative estimate. It results in substantial computational cost savings over the DB method, more so for data sets with more OTUs. In particular, the parametric version of CVB is recommended for implementation, where the bootstrap samples are drawn from the parametric distribution determined from the model at the value of the MLE.

It is important to note that the above bootstrap methods for estimating the variability of the MLE for evolutionary parameters can be incorporated into the standard bootstrap analysis commonly performed after ML estimation without much additional computational time costs and the implementation can be easily carried out. It is common to perform a bootstrap analysis to assess support for clades in the resulting ML tree topology. Depending on the approaches that are taken, either the DB or CVB method can be added. If a bootstrap analysis is to be performed with evolutionary
parameters fixed at their ML values from the original data set, the so generated bootstrap samples can be used to implement the CVB method. For each bootstrap sample, the ML tree under the fixed parameter values (MLE of the original data set) is already determined from the standard bootstrap analysis for topology. The additional step is to carry out the CVB method, that is to find the evolutionary parameter value that maximizes the likelihood under the ML tree, for which the computational cost is minimal. On the other hand, if the standard bootstrap approach already finds the global MLE of tree topology and evolutionary parameters for each bootstrap sample, the DB method can be used to report the variability of ML estimates among these bootstrap samples.

Moreover, although the simulations were demonstrated only for parameters $\mu$ and $K$, the method proposed is not limited to them. Theorem 3.2.1 laid the foundation for the asymptotic behavior of the MLE for all of the substitution model parameters and the branch lengths vector. If a phylogenetic analysis package can handle simultaneous estimation of the nucleotide equilibrium frequency parameters, the bootstrap methods described can handily apply as well.

The maximum likelihood estimation in the phylogenetic analysis context has certain issues, as does other ML estimation in an unbounded parameter space. However, the likelihood surface is perhaps more complex than for most other applications.

For a particular data set, the maximum of the likelihood surface can be attained for evolutionary parameters at the boundary of parameter space, that is, the MLE for $\mu$ and $K$ may be $0$ or $\infty$. A simple example is to consider a data set with identical nucleotide sequences at all the OTUs, which would yield the MLE of $\mu$ being $0$. 

131
However, such data sets do not contain much phylogenetic information, and are not of practical interest.

In connection with this observation, the true variance of the ML estimator of $\mu$ and $K$ can be $\infty$, since there can be a positive probability of drawing a data set with a finite number of sites which results in an infinite ML estimate of $\mu$ and/or $K$. This was pointed out by a reviewer of [125], Jeff Thorne, as well. The large-sample property established in Chapters 2 and 3 pertains to the asymptotic variance-covariance matrix of the limiting distribution of the MLE for evolutionary parameters, and branch lengths vector, if applicable. Fortunately, these results imply that the aforementioned positive probability for a finite number of sites approaches zero as the number of sites increases. The aim of the estimation for variability is to provide a practical guide on the uncertainty about the estimate. When tree topology and branch lengths vector are considered known, the MCS method for expected Fisher information could be employed to approximate the asymptotic variance of evolutionary parameters. When they are unknown and are estimated, bootstrap methods proposed can provide an assessment of the variability of the estimate. There are two approaches one could take for dealing with the interpretation of the variability. We could limit the space of $\mu$ and $K$ into a bounded rectangle with very large upper limits for $\mu$ and $K$. The parameter estimates and those from bootstrap samples are truncated. Another approach is to use a transformation of $\mu$ and $K$ so that the parameter space is bounded. It is important to note that this infinite true variance occurs for other applications of ML estimation with an unbounded parameter space as well.

As in other applications, multiple maxima for the likelihood surface can occur for particular data sets in phylogenetic analysis [19, 97]. Theorems 2.2.2 and 3.2.1
that were proved in Chapters 2 and 3 entail that the probability of having multiple maxima approaches 0 as the number of sites increases. In the case of a data set with multiple maxima, it is expected that the variability estimate from bootstrap methods becomes large. If the uncertainty about the MLE is markedly large for a data set, that may be an indicator of multiple local maxima with close likelihood values and that caution needs to be used. From simulation, it is observed that the nonparametric bootstrap samples by resampling columns of the data occasionally produce multiple maxima, but much more frequently than the parametric bootstrap method. It is one of the reasons why the parametric bootstrap is recommended over the nonparametric counterpart.

Theorems 2.2.2 and 3.2.1 guarantee that the distribution of the MLE for a data set with a large number of sites is close to the multivariate normal distribution. Nonetheless, as large as a genome is, it never has an infinite number of nucleotides. Based on Figures 2.4 through 2.11 for Simulations 2 and 3, the top panels show that the distributions of the MLE of $\mu$ and $K$ reasonably resemble normal distributions even for the simulated data sets for 14 sequences with 231 sites. However, as expected with the absence of a right boundary for the parameter space, their distributions, especially for $K$, appear to be skewed to the left. It suggests a transformation, perhaps logarithm or a negative power function, on the parameters will likely improve the normal approximation and also the performance of estimators of variance, both in the case of asymptotic variance assuming the tree is known and the bootstrap method with the tree simultaneously estimated.

Yang’s PAML [134] is the only phylogenetic analysis package known to give an estimate of the variability for evolutionary parameters. Though the asymptotic efficiency
was never proved, PAML inverts the observed Fisher information matrix (curvature method) and reports the s.e. of evolutionary parameter MLE as well as that for estimated branch lengths vector. Through personal communication, Yang had reported that the difference approximation was used to obtain the observed Fisher information and it “might not always be reliable”. However, it is worth reiterating that with very few exceptions, the ML tree topology is not necessarily the true topology. The computation of Fisher information has to make that assumption, and the resulting estimates of the variability suffer from downward bias as it ignores the uncertainty with estimation of topology. What is worse is that we cannot test how close the ML topology is to the true topology and how severe the underestimation is for a data set. The bootstrap methods are recommended for the situation for which the tree topology and branch lengths vector need to be estimated. In the rare situation where the tree topology can be assumed to be known, the MCS method via sampling from the parametric distribution with the parameter fixed at the MLE and approximating the expected Fisher information is likely a superior method to the curvature method.

MCMC methods have also been widely used for phylogenetic analysis [76, 82, 137, 63], and the posterior high-probability region provides a credible set for parameters of interest in the Bayesian paradigm. Although it may require more computation time, the approach can get information on the entire posterior distribution as compared to just the variability estimate from the bootstrap approach. But certain common problems with MCMC will need to be dealt with, such as the diagnostics of convergence and how to implement the prior.
CHAPTER 5

FUTURE DIRECTIONS

In this thesis, the large-sample distributional property for the joint MLE of tree topology, branch lengths vector and evolutionary parameters, as the number of sites approach infinity, is established. A practical bootstrap method was proposed to estimate the variability of the ML estimate of evolutionary parameters. The estimation of evolutionary history for nucleotide sequences and proteins is an exciting area where statistics has been and will continue to be relied upon. Here are some open problems for further research. They are organized into two types – those concerning the phylogenetic tree construction alone and those for simultaneous alignment and tree inference.

5.1 Tree Construction Related Problems

A major difficulty, as well as a fascinating aspect, for the subject of phylogenetic analysis is the complex space and intertwining nature of tree topology and the branch lengths vector. Tree topology forms a discrete space with a very large cardinality. On top of that, for each tree topology, there is a set of congruent branch lengths vectors that is a subset of a high-dimensional Euclidean space. Tree topology in itself does not convey all the information regarding evolutionary history, e.g., a tree
with a different phylogeny can have a higher likelihood than a tree with the same phylogeny as the ML tree. An ambitious proposal is to formulate a metric on this complex space, which may be adaptive to the observed data. This metric is defined on the space of all possible unrooted binary trees and star phylogenies. The desirable property for this metric is such that not only is it a useful tool for summarizing the trees, but also as a way of defining a “confidence region” of the tree space. Such a metric will allow a unified approach on the complex space of tree topology and branch lengths and the large-sample theory that was established in this thesis can then be generalized to this tree space. This metric can address the issue of estimating the variability for tree topology and branch lengths. Somewhat connected to this idea, it is desirable to quantify the phylogenetic information contained in a data set, and therefore, the uncertainty associated with the tree estimation could be described. Even without a viable metric, this may still be possible. A data set must have an appropriate level of phylogenetic information for the analysis to be meaningful. At the two ends of the spectrum are the situation where the sequences at the external nodes are simply a random sample from the stationary distribution, and the situation where the nucleotide sequences at the external nodes barely have differences, neither of which contains much phylogenetic information.

In the case where the assumption of independence and identical distribution across sites is not reasonable for a nucleotide data set, more theoretical research needs to be done. In addition, practical methods need to be devised to evaluate these assumptions.
5.2 Simultaneous Alignment and Phylogeny Construction

Another assumption we have made so far with the data set is that the alignment of data is known. However, for nucleotide or amino acid data that are generated by sequencing techniques, all that is presented is the contiguous characters without the knowledge of the positional homology among the sequences under study. It is of interest to find the alignment of homologous sequences, i.e., to recover the evolutionary correspondence of nucleotides (or amino acids) in one sequence with those in other sequences. To align sequences, one adds possible gaps to sequences to make them have equal lengths, and lines them into a matrix. Characters, which may be nucleotides (amino acids) or added gap, in each column display positional homology. In the columns with gaps, insertion or deletion events are inferred to have occurred. Columns with only gaps are not allowed.

<table>
<thead>
<tr>
<th>sequences</th>
</tr>
</thead>
<tbody>
<tr>
<td>seq1</td>
</tr>
<tr>
<td>seq2</td>
</tr>
<tr>
<td>seq3</td>
</tr>
</tbody>
</table>

Table 5.1: Example for three hypothetical unaligned nucleotide sequences.

A hypothetical data set of nucleotide sequences with different lengths of three homologous taxa is displayed in Table 5.1. Figure 5.1 shows three possible alignments along with three distinct phylogenies for this data.

The number of distinct alignments grows even more rapidly than that for phylogenies. There are \( \binom{n+m}{n} \) ways to align two sequences of \( m \) and \( n \) nucleotides long, respectively. For \( N \) sequences with length \( l_1, l_2, \ldots, l_N \), the number of alignments is much bigger than the multinomial coefficient, \( \binom{\sum_{i=1}^{N} l_i}{l_1 l_2 \ldots l_N} \). The number of
the distinct combinations of tree and alignment is the product of the number of distinct trees of given taxa and the number of distinct alignments. For example, there are 93 distinct tree–alignment combinations for 3 sequences with only 1, 1, and 2 nucleotides, respectively.

5.2.1 Background

There is a plethora of literature on the subject of alignment in computational biology. For global alignment of two sequences under a pre-specified gap penalty function, dynamic programming type methods [8, 87] have been applied and are guaranteed to find the optimal alignment(s). The limitation is that the gap penalty function has to be given a priori. Attempts were made in the literature to put the in/del and point mutation into a probabilistic model-based paradigm [10, 117, 118], but computational costs of more realistic models are great and cannot be implemented.

Realistically, most of the alignment situations one may encounter are the alignment of multiple sequences when doing phylogenetic tree construction and searching a sequence database. Multi-dimensional dynamic programming, a direct generalization of dynamic programming algorithm for two sequences, requires both a prohibitive
amount of computing time and memory and is not practical for the alignment of more
than four sequences. Heuristic methods have been proposed [42, 7], but they cannot
guarantee finding the optimal alignment under a pre-specified penalty function. Re-
cently, Hidden Markov Models (HMM), first applied in the speech recognition field in
the 1970s [93], have become the most popular method, especially for protein sequence
data, in multiple sequence alignment [68, 26]. It assumes that there is an underlying
unobservable Markov chain, while there are parameters that govern the link between
unobservable states and the observable characters. Also the Markov chain is char-
acterized by a set of parameters for transition probabilities between states. Like the
approach for machine learning, the parameters can be estimated well with a large
training set, and in turn describe the alignment. HMM works more or less like a
black box, but it provides a practical way to align sequences, especially for search-
ing a protein database to infer the function of a new protein sequence. However, it
ignores the fact that some sequences are closer in evolutionary history than others.
The choice of the number of states in the underlying Markov chain is an art, and the
estimation of the many parameters can easily get stuck in a local optimum. Another
approach in this alignment context is the Gibbs sampler type of method to find reg-
ulatory motifs in unaligned DNA sequences by J. Liu and his coworkers [73, 79, 80].
The problem is formulated as follows: each sequence contains highly conserved re-
gions(s), at which the residuals are much more similar across different sequences than
at other regions. In single-motif situation, residues within the motif are modeled by a
product multinomial distribution, and those outside of the motif are treated as i.i.d.
observations from a common multinomial distribution. By applying a Dirichlet prior
on the multinomial parameters, a Gibbs sampler is constructed to approximate the
posterior probability of the motif starting with a specific position. It can be viewed as a special case of HMM. However, the above approaches all suffer a common setback – they ignore the phylogenetic structure of the sequences under study. One would expect the sequences that diverge later in evolutionary history to be similar both in terms of the residuals at the homologous sites and also in terms of similar alignment. It is not appropriate to align all the sequences based on the same criteria regardless of how far apart sequences were in evolution.

Phylogenetic tree construction and multiple sequence alignment are two integral parts of a single problem. Sequences are related by the underlying phylogenetic tree structure, and on the other hand, the position homology is not clear in observed homologous sequences. Almost all the current tree building methods assume the alignments are given, whereas the methods to construct multiple sequence alignment either ignore the underlying tree structure and assume independence of the sequences, or assume that the tree is given and apply some ad hoc weighting scheme. This circular argument has been pointed out in many articles [104, 20, 73]. ClustalW [57], a widely used alignment package, uses a progressive alignment approach: a “guide tree” is inferred by distance-based method using the pairwise alignment penalty scores, and following this tree, neighboring sequences on the tree are then aligned with each other. Also of note is that there should be a coherent system for both alignment and phylogenetic analysis [59, 115]. There had been two methods of simultaneously aligning sequences and finding a plausible phylogeny for them in the literature. Sankoff & Cedergren [103] and Hein [55] both attempted to find an optimal alignment given a tree, with the first implementing a linear gap penalty and the latter an affine gap penalty. Both methods are parsimony-type algorithms, trying to minimize the sum of
total residue change in the tree. They both require a search of all the tree topologies to find the overall optimum, which cannot be practically used.

5.2.2 A Proposal for Simultaneous Tree Inference and Sequence Alignment

A new Metropolis-Hastings algorithm will be proposed to handle simultaneous sequence alignment and phylogenetic tree construction. The general Markov chain Monte Carlo method will be first briefly described, then it is followed by the evolutionary model of modeling point substitution and point in/del. Notations and the moving strategy for the new algorithm will then be introduced.

Markov Chain Monte Carlo

Markov chain Monte Carlo (MCMC) [119] is a simulation method that generates a Markov chain with the stationary distribution resulting in the desired distribution. Quite often, the probability distribution of interest is not of closed form, and MCMC provides a good practical remedy by generating a dependent sample.

Suppose we would like to find certain statistics of a probability distribution \( \pi(x), x \in \mathcal{X} \), or even sometimes \( \pi(x) \) itself. Let the \( \sigma \)-field on \( \mathcal{X} \) be \( \mathcal{B} \), and \( \pi \) is a probability density with respect to a measure \( \mu \). Let \( P(\cdot, \cdot) \) on \( \mathcal{X} \times \mathcal{B} \) be a transition probability function, such that

(i) \( P(x, \cdot) \) is a probability measure on \( (\mathcal{X}, \mathcal{B}) \), for each \( x \in \mathcal{X} \);

(ii) \( P(\cdot, C) \) is measurable with respect to \( (\mathcal{X}, \mathcal{B}) \), for each measurable set \( C \in \mathcal{B} \).

A Markov chain \( \{X_n; n \geq 1\} \) is generated from \( P \) as the transition probability function. If \( P \) satisfies certain conditions, the resulting Markov chain is ergodic, or the distribution of \( X_n \) converges to \( \pi \).
The Metropolis-Hastings sampler is a method to construct a transition probability function $P(\ldots)$. Let $q(\cdot, \cdot)$ be defined on $\mathcal{X} \times \mathcal{X}$, where $q(x, \cdot)$ is a probability density with respect to $\mu$ for each $x \in \mathcal{X}$. For any $x, y \in \mathcal{X}$, define the acceptance function $\alpha(x, y)$ by

$$
\alpha(x, y) = \begin{cases} 
\min\left\{ \frac{\pi(y)q(y,x)}{\pi(x)q(x,y)}, 1 \right\}, & \text{if } \pi(x)q(x,y) \geq 0; \\
1, & \text{if } \pi(x)q(x,y) = 0.
\end{cases}
$$

The Metropolis-Hastings sampler is generated as follows: suppose that the chain is currently at a state $X_n = x$, a candidate state $y$ is drawn according to the transition kernel density $q(x, \cdot)$. With probability $\alpha(x, y)$, the chain moves to the candidate state, $X_{n+1} = y$. With the remaining probability $1 - \alpha(x, y)$, the chain stays at $x$, $X_{n+1} = x$. Under certain conditions, the distribution of the chain generated by the above sampler converges to $\pi$. Note that the acceptance probability depends only on the ratio of $\pi(x)$ and $\pi(y)$, hence it is particularly useful for sampling from a distribution whose normalizing constant is hard to compute.

**Evolutionary Model**

In order to facilitate the likelihood calculation, we will specify the evolutionary model of point substitution and point in/del. We model the evolution of the homologous nucleotide sequences with the following assumptions:

1. Mutations at each site (each column of the true alignment) occur according to a homogeneous Poisson process.

2. Mutations at different sites are independent.
We model the evolutionary process as a continuous-time finite-state Markov process. The state space is over the alphabet $\mathcal{A} = \{ A, C, G, T, - \}$, where 
"-
 represents the gap. The infinitesimal transition matrix is given by

$$
\begin{pmatrix}
-(3 + \gamma)\alpha & \alpha & \alpha & \alpha & \gamma\alpha \\
\alpha & -(3 + \gamma)\alpha & \alpha & \alpha & \gamma\alpha \\
\alpha & \alpha & -(3 + \gamma)\alpha & \alpha & \gamma\alpha \\
\gamma\alpha & \gamma\alpha & \gamma\alpha & \gamma\alpha & \gamma\alpha \\
\gamma\alpha & \gamma\alpha & \gamma\alpha & \gamma\alpha & 4\gamma\alpha
\end{pmatrix}
$$

where parameter $\alpha$ reflects the rate of substitution, and $\gamma$ governs the ratio of the rate of in/del events to the substitution events. The transition probabilities from letter $i$ to $j$ in time $t$ is given by

$$
P_{ij}(t) = \begin{cases}
\frac{1}{5} + \frac{3}{4}e^{-(1-\gamma)\alpha t} + \frac{1}{20}e^{5\gamma\alpha t}, & \text{if } i = j \in \{ A, C, G, T \}; \\
\frac{1}{5} - \frac{3}{4}e^{-(1-\gamma)\alpha t} + \frac{1}{20}e^{5\gamma\alpha t}, & \text{if } i, j \in \{ A, C, G, T \}, i \neq j; \\
\frac{1}{5} - \frac{1}{5}e^{-5\gamma\alpha t}, & \text{if } i \in \{ A, C, G, T \}, j = - \text{ or vice versa}; \\
\frac{1}{5} + \frac{4}{5}e^{-5\gamma\alpha t}, & \text{if } i = j = -
\end{cases}
$$

This model implies that the in/del events occur one nucleotide at a time.

**Problem Formulation**

Suppose that we have the $N$ homologous nucleotide sequences. We adopt a Metropolis-Hastings algorithm to study the conditional distribution of the combination of tree and alignment given the observed data set. The notations of the components of the study object are as follows:

$T$: tree topology;

$t$: vector of branch length of the tree;
**v:** the internal node sequences with possible padded gaps;

**b:** the external node sequences with possible padded gaps.

The observed sequences \( b' \) are the aligned sequences of \( b \) with all the gaps removed. Define the state space \( \Omega \), where each state \( (T, t, v, b) \) is the collection of the phylogenetic tree topology, branch length, the internal node sequences and external node sequences. If we are only interested in recovering the phylogenetic tree topology and the alignment, the distribution of interest is

\[
P(T, b|b') = \frac{P(T, b, b')}{P(b')}.\]

In the implementation of the Metropolis-Hastings sampler that follows, we construct the Markov chain in the state space \( \Omega \), and update all of the four components of the object. One reason for carrying more variables in the updating scheme is that difference in the likelihood function between candidate state and current state becomes easier to compute. Another reason is that it allows the alignment to be updated based on branching time and avoids drawing the candidate state from low probability regions too often.

We wish to sample from the distribution:

\[
P(T, t, v, b|b') = \frac{P(T, t, v, b)}{P(b')},
\]

noting that the information of \( b' \) is superfluous given \( b \). The distribution of a totally specified object is given by

\[
P(T, t, v, b) = \pi(T) \pi(t|T) \pi(v_0) \prod_{u: \text{parent node of } w} P(w|u),
\]

where \( v_0 \) is the root node. Following Li, Pearl and Doss [76], the prior on the topology \( T \) and the branching time \( t|T \) is chosen to be an uninformative prior. The prior on
$T$ is a uniform prior among all the distinct rooted topologies for the $N$ OTUs. The prior on the vector of branching time $t$ given the topology $T$ is derived from the order statistics of a uniform distribution. The prior on the root node is the product of the density function of the length of the root node sequence or the alignment and the stationary distribution density of each letter \{A, C, G, T, \_\} at each position. For each site (each column of the alignment), the probability of changes in nucleotides along each branch is given by the transition probability induced from the infinitesimal transition probability matrix above.

**Moving Strategy**

Given a current state $x \in \Omega$, a candidate state is generated as outlined by the following steps.

1. Draw a random internal node, but not the root node. Use the local rearrangement strategy as illustrated in Figure 1.2 to draw one out of three possible tree topologies based on the current tree.

2. Draw the time associated with the target node.

3. Use a dynamic programming idea to align the two child nodes of the current target node. Store a probability vector (profile) for each aligned site of the target node.

4. Align the target node with its sibling. Draw the nucleotide sequence at the target node based on the stored profile probabilities of the target node and the aligned sequence of its sibling node. Store the profile probability vector for each aligned site of the parent node.
5. Repeat Step 4 by substituting the parent node into the role of target node. This bottom-up updating process does not stop until the target node becomes the root node.

Step 1 updates the topology using the local rearrangement strategy, which was illustrated in Figure 1.2 [69, 76]. After the target node is drawn, one out of three nodes (two child nodes and the sibling node of the target) is randomly drawn to be the sibling node of the target in the candidate topology. The remaining two nodes are attached to the target node as the new child nodes. All the subtrees starting from those three nodes remain the same. The local rearrangement strategy provides a good compromise between mixing up well in the space of topologies and not visiting the trees in the low probability region too often. In addition, the calculation of likelihood for the candidate state can make use of portions of the likelihood function calculated for the current state, since the likelihood for the subtrees not involved in the rearrangement remains the same.

Step 2 updates the branching time vector. The recipe is that if the topology of the candidate state is the same as that of the current state, a Beta distributed time vector is drawn from the interval between time of the parent and the earlier branching time of the two daughter nodes. If the topology is different, a uniformly distributed time is drawn from the proper time interval.

In this proposed Metropolis-Hastings sampler, the multiple alignment is formed by progressively aligning two sequences at a time. At each node to be updated, we generate the nucleotide sequence after the alignment of its two daughters is drawn. Therefore, each subtree from a generic internal node cannot have an all-gap column when the alignment is constructed from the bottom up. When two daughter nodes
Figure 5.2: Illustration of the dynamic programming method to draw a random pairwise alignment.

are aligned, a number of gaps may be added to one or both nodes, and the gap(s) will be carried to each descendent node of the daughter nodes at the same positions. After the two daughter nodes of the root node are aligned, a top-down pass can trace all the sequence gaps, and thus specify a multiple alignment.

In Step 3, aligning two daughter node sequences is performed by a dynamic programming scheme, similar to the one in Churchill [21]. Suppose that we have two sequences $A = a_1 a_2 \cdots a_{n_A}$ and $B = b_1 b_2 \cdots b_{n_B}$. We form a grid of vertices, for which each edge represents a character, either the observed character or unobserved gap (Figure 5.2). This two-dimensional grid is referenced by row index $i = 0, 1, \ldots, n_B$, and column index $j = 0, 1, \ldots, n_A$. An alignment can be thought of as a path from the upper left corner to the lower right corner, with each arc being one of the three moves: south (\downarrow), east (→), southeast (↘). Thus an alignment can be represented by a series of arcs:

$$\alpha = \alpha_1 \alpha_2 \cdots \alpha_n,$$
where $n$ is the length of the alignment and 

$$
\alpha = \begin{cases} 
\rightarrow, & \text{adding a gap in B;} \\
\downarrow, & \text{adding a gap in A;} \\
\wedge, & \text{matching the two bases in A and B.}
\end{cases}
$$

The dynamic programming is done in two passes: forward pass and backward pass. In the forward pass, we calculate the unconditional probability, $P(A_i, B_j)$, of aligning two subsequence $A_i = a_1 \cdots a_i$ and $B_j = b_1 \cdots b_j$ by summing the three terms: $P(A_{i-1}, B_j) P(a_i, \_)$, $P(A_i, B_{j-1}) P(\_ b_j)$ and $P(A_{i-1}, B_{j-1}) P(a_i, b_j)$. The three terms represent aligning $A_i$ with a gap, a gap with $B_j$, and $A_i$ with $B_j$, respectively. This determines the central recursion in the dynamic programming. We store a conditional probability vector, whose three entries are the ratios of the above three terms to the total $P(A_i, B_j)$. The boundary condition is the obvious one. When the forward pass reaches the lower right corner, we start the backward pass. At each vertex, we sample an arc from the trinomial distribution specified by the three stored proportions corresponding to the three arcs entering the vertex. This traceback ends at the upper left corner, and we are left with a random alignment of the two sequences.

After a pair of daughter nodes’ sequences are aligned, we store a profile probability vector at the parent node for each site of the alignment. The profile vector at each site saves five entries. Each entry corresponds to the probability of nucleotide at the parent node at the site being one of the five letters in $\mathcal{A}$. The probabilities sum up to 1, and are consistent with the conditional distribution given the two daughter node nucleotides at the same site and the branch lengths on the two branches under the evolutionary model.

The alignment of the profile vector of a node with its sibling (a sequence) is straightforward extension of the above dynamic programming scheme. In this case,
each site in the node has only a profile vector, and the calculation of the probability of
aligning it with a letter in another sequence is done by a weighted average, weighted
by the probabilities in the profile vector. As in Step 4, after the target node (which
only has a profile vector) is aligned with its sibling, a letter from \{A, C, G, T, \_\} is
drawn at each aligned site from the conditional distribution given its sibling node and
the profile vectors based on the evolutionary model. Once the nucleotide sequence at
the target node is generated, a profile probability vector for the parent node is formed
in the same way as in step 3. Then, we replace the role of target by its parent node,
and repeat step 4, until the root is reached.

5.2.3 Results

Two numerical examples were tested for the proposed Metropolis-Hastings sam-
pler.

The first example is concerning three short sequences, for which the theoretical
probability of each alignment-phylogeny pair under model parameters \(\mu\) and \(\gamma\) can
be calculated analytically. The sequences are A, C, and AC. We fix the parameter
values at \(\mu = 0.05\), and \(\gamma = 0.08\). The likelihood of each possible combinations of
tree-alignment combination is calculated by the pruning algorithm that was described
in Section 1.5. There is a total of 12 tree-alignment combinations for alignments of
length 2, 45 of length 3 and 36 of length 4. We adopt parenthesis notion to represent
tree topologies. For example, \(((A, C), (A, C))\) represents the tree which has sequences
with nucleotides A, and C branch off later than their parent node and sequence AC do.
In addition, the alignment is specified by the letters in the first column A, \_ A; and
those in the second column: \_ C, C. Table 5.2 lists out the conditional probabilities of
top eight tree-alignments, all are of length 2, given the observed compact nucleotide sequences. The next four combinations with largest conditional probabilities are \(((\_A, \_C), AC)\) with conditional probability of 0.006, \(((\_A, C\_), AC)\) with probability $7.97 \times 10^{-5}$, \((\_A, AC), C\_\) with probability $1.02 \times 10^{-4}$, and \(((C\_ AC), A)\) with probability $1.02 \times 10^{-4}$, all of which are of length 2. The 45 tree-alignments combinations with alignment of length 3 all have conditional probabilities less than $2 \times 10^{-5}$, and all the combinations with length 4 have conditional probabilities in the order of $10^{-9}$.

The results of simulation are summarized into the cell counts of nine cells: the eight tree-alignments with highest conditional probabilities and a cell combining the others.

The initial starting point of the tree topology was \(((A, C), AC)\). The first 2000 burn-in samples of the chain were discarded. Then the chain was run for another 1000000 steps. Every 100-th state was sampled and kept for the analysis. The resulting 10000 samples were categorized into the nine cells as described earlier, and the expected counts and observed counts are shown in Table 5.2.

<table>
<thead>
<tr>
<th>Tree-alignement</th>
<th>Conditional probabilities</th>
<th>expected cell count ((m_i))</th>
<th>observed cell count ((n_i))</th>
<th>((n_i - m_i)^2/m_i)</th>
</tr>
</thead>
<tbody>
<tr>
<td>1. (((A_ AC), A_))</td>
<td>0.0133</td>
<td>132.59</td>
<td>139</td>
<td>0.310</td>
</tr>
<tr>
<td>2. (((C_ AC), A_))</td>
<td>0.0133</td>
<td>132.59</td>
<td>137</td>
<td>0.147</td>
</tr>
<tr>
<td>3. (((A_ C_), AC))</td>
<td>0.1958</td>
<td>1958.19</td>
<td>1991</td>
<td>0.550</td>
</tr>
<tr>
<td>4. (((A_ AC), C_))</td>
<td>0.1919</td>
<td>1919.12</td>
<td>1947</td>
<td>0.405</td>
</tr>
<tr>
<td>5. (((C_ AC), A_))</td>
<td>0.0958</td>
<td>957.65</td>
<td>944</td>
<td>0.195</td>
</tr>
<tr>
<td>6. (((_A, C_), AC))</td>
<td>0.1958</td>
<td>1958.19</td>
<td>1932</td>
<td>0.350</td>
</tr>
<tr>
<td>7. (((_A, AC), _C))</td>
<td>0.0958</td>
<td>957.65</td>
<td>951</td>
<td>0.046</td>
</tr>
<tr>
<td>8. (((_C, AC), _A))</td>
<td>0.1919</td>
<td>1919.12</td>
<td>1892</td>
<td>0.383</td>
</tr>
<tr>
<td>9. Others</td>
<td>0.0065</td>
<td>64.91</td>
<td>67</td>
<td>0.067</td>
</tr>
</tbody>
</table>

Table 5.2: Summary of the expected vs. observed cell counts for the example with three observed sequences A, C and AC. The eight most likely tree-alignments and the category combining all others are presented.
In the first eight cells where the tree-alignment has larger conditional probability, the expected cell count and observed cell count are close, and the contribution to the goodness-of-fit χ² statistic is small. In the last cell that combines the relatively rare tree-alignment combinations, the difference between the observed and expected counts is also small. The goodness-of-fit χ² statistic for the above nine-cell multinomial is 2.45, and the associated p-value is 0.964 for a chi-square with 8 degrees of freedom.

| seq1 | TGTCTCAGCATACCTGCCTTTGCAAGCAACGATTCTTATAGGAAATACGAGCGTCCAG - G |
| seq2 | TGTCTCACCATCACTGCTT-TGCACGCAACGATCTTAGGAAACGCTTCGCCAG - G |
| seq3 | TGTCTCTGATTTGCACTTGAATTTTTTAGGAGAATTCGAGCCAG-

Six aligned sequences of 60 sites displaying gaps

| seq1 | TGTCTCAGCATACCTGCCTTTGCAAGCAACGATTCTTATAGGAAATACGAGCGTCCAGG |
| seq2 | TGTCTCACCATCACTGCTT-GTCACGCAACGATCTTAGGAAACGCTTCGCCAGG |
| seq3 | TGTCTCTGATTTGCACTTGAATTTTTTAGGAGAATTCGAGCCAGA |
| seq4 | TGTCTCTGATTTGCACTTGAATTTTTTAGGAGAATTCGAGCCAGA |
| seq5 | TGTCTCTGATTTGCACTTGAATTTTTTAGGAGAATTCGAGCCAGA |
| seq6 | TGTCTCTGATTTGCACTTGAATTTTTTAGGAGAATTCGAGCCAGA |

Resulting unaligned sequences by removing gaps from sequences in the above tabulation

Another example that was tested is a data set generated from a known tree and known parameter values. The underlying true tree was the asymmetric tree from the sequences 1 through 6, (((seq1, seq2), seq3), seq4), seq5), seq6), with branch
lengths being equal for all pairs of parent-daughter internal nodes. The parameter values were set at $\mu = 0.08$, and $\gamma = 0.08$. At first, 60 nucleotides were drawn with equal probability of being \{A, C, G, T, -\} to form the nucleotide sequence at the root node. Following each branch from the top to bottom in the pre-specified tree, we generated nucleotide sequences at each node. The sites with only gaps in the six external node sequences were removed. The above tabulations show the aligned sequences with gaps and the unaligned sequences with gaps removed as if they were the observed sequences.

We fixed $\mu$ and $\gamma$ at the underlying true values when carrying out the MCMC runs. Two different starting trees were tested. In the first experiment, the initial tree was the true underlying tree. 5000 burn-in runs were discarded, and an additional 50000 states in the chain were recorded. It turned out that 69.5\% of the trees had the same topology as that of the underlying true tree, and 96.92\% of the states had the true alignment (without the all-gap columns). The chain accepted 6172 times out of 16840 cases where the candidate state had the same topology as the current state; while it accepted 4360 times out of 33160 cases where the candidate state had a different topology.

In the second experiment, we started the tree at the opposite asymmetric tree: ((((((seq5, seq6), seq4), seq3), seq2), seq1). The same number of burn-ins and the samples were taken. 74.4\% of the trees had the same topology as the underlying true tree, and 95.19\% had the alignment with 60 sites. The chain accepted 6188 times out of 16644 cases where the topologies remained the same for the current state and the candidate state, yet 3933 times out of 33356 cases where they differed.
5.2.4 Discussion

This MCMC method is the first attempt that puts both multiple alignment and phylogenetic tree estimation into the likelihood framework with a parametric model explicitly modeling the evolutionary process. When an alignment is inferred according to a specific tree topology, the subsequent inferred phylogenetic tree will be in favor of that topology (Thorne & Kishino [115]). By considering both alignment and phylogenetic tree altogether, we avoid breaking an integral estimation problem down into multiple alignment and tree reconstruction separately.

The best alignment, even in the pairwise case, is rarely the true alignment from the evolution, since there is a huge number of variants of the same alignment with nearly the same probability. For example, the best alignment of human alpha globin to leghaemoglobin accounts for only $4.6 \times 10^{-6}$ of the posterior probability of being the correct alignment, (Durbin et al. [26]), where each globin has only 28 or so amino acids. Therefore, it is well-recognized that finding the single best alignment may not be the best problem formulation for the sequence alignment problems. This MCMC method samples from the posterior distribution, and automatically gives the sub-optimal alignments around a mode of alignment in the likelihood surface.

With such a complex object of alignment and phylogenetic tree considered together, it is often the case where there are many modes in the likelihood surface. Different alignments support different inferred trees. It is my hope that this type of MCMC method can simultaneously examine the modes, compared to the method proposed in the literature which derive only one mode but not necessarily the global maximum due to the heuristic nature of the searching algorithm. The confidence to
each pair of alignment and phylogenetic tree is handily summarized by the posterior probability.

This current method has had success in small examples. However, when a larger data set with ten or so sequences and about 50 nucleotides was tested, the method did not appear to be a viable approach. The proposed state gets accepted too infrequently because as the complexity of the problem increases, the method could not sample other high-likelihood areas without an extreme computational burden. There is too big a state space to traverse. The proposal of the new alignment is not coupled with the local rearrangement strategy of moving in the topology space. Although the idea shows promise, a new method to draw proposed state that better handles the movement in the topology space and alignment space needs to be sought. Preferably, a clever moving strategy that couples the move in the tree space and in the alignment space can be found. Since alignment space cardinality increases much faster with the number of residuals and also its nature renders that changes of alignment in a local area affects other areas of the alignment, it does not appear to be paired with the local rearrangement movement for the tree space well. Amino acid data sets in general have shorter sequences, and the method I proposed may perform better in that setting. However, the large number of character states presents a challenge for determining the substitution matrix.

Another shortcoming with the proposed method is that the model for nucleotides and gaps is not realistic. This model can be easily extended to a more realistic substitution model between the nucleotides, but the assumption of the single-base-long gaps deviates from the real in/del generation mechanism. Development of more realistic stochastic models of gaps and substitution is needed. However, it presents
problems as the model complexity grows. The first is computational and it may be remedied by the rapidly improving computer hardware. The second is philosophical in that no model is correct. Careful analysis of the model robustness can be used to address this issue with extensive simulation studies. In addition, a complex model also presents great difficulties with limited amount of data to estimate the model parameters and the object of interest.
In closing, I would like to quote Cavalli-Sforza’s remark in the discussion of Edwards’s 1970 *JRSSB* paper [27]

“I am very pleased to see that the problem offers sufficient challenge to statisticians.”

Three decades later, this area of studying the phylogenetic analysis and alignment for the nucleotide sequences and amino acid sequences has continued to present challenges and interesting problems for statistical applications. Recently developed computational intensive methods, such as bootstrap, MCMC, and machine learning, have all found its share in this area. It is expected that other new statistical tools will continue to be applied and contributed to this exciting research subject.
BIBLIOGRAPHY


166


