INFORMATION TO USERS

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

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

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

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

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

DISSERTATION

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

By

Xiong Hu, B. S., M. S.

* * * * *

The Ohio State University

1998

Dissertation Committee:
Professor Dennis K. Pearl, Adviser
Professor Douglas E. Critchlow
Professor Elizabeth A. Stasny

Approved by
Adviser
Graduate Program in Biostatistics
The ratio of the synonymous substitution rate ($K_S$) to the nonsynonymous substitution rate ($K_A$) is a measure of biological functionality that is commonly used by molecular biologists. In this dissertation, a simple method is proposed to estimate $K_S$, $K_A$, the ratio $K_S/K_A$, and the variances of these estimators. The estimators have been developed based on both Jukes and Cantor's one-parameter model, and Kimura's two-parameter model. Recombination is a major source of genetic diversity. One way to detect recombination is in the context of phylogenetic analysis. A likelihood-based method is developed to detect any differences between two underlying phylogenetic trees by using nucleotide sequences. To ensure the independence between the inferred tree and the nucleotide site being evaluated, a leave-one-out technique is employed when computing the maximum likelihood value of each polymorphic site. Both of these two methodologic developments are applied to problems in HIV disease.
ACKNOWLEDGMENTS

I wish to thank Professor Dennis Pearl for guiding me through this dissertation, always encouraging me and always patient. I wish to thank Professor Douglas Critchlow for his help, insight and encouragement. I also wish to thank Professor Elizabeth Stasny whose comments on this dissertation were very helpful. I wish to thank my parents and grandmother for their love all these years. Finally, I wish to thank my wife, Ling, for being there through this entire ordeal when she herself has to work very hard for her dissertation. The love and support she has given me these past years are beyond words.
VITA

June 25, 1968 ......................................................... Born - Guangzhou, P.R. China

June 1990 .............................................................. B.S. Computational Mathematics, Peking University, Beijing

March 1995 ............................................................. M.S. Statistics, The Ohio State University, Columbus, Ohio

FIELDS OF STUDY

Major Field: Biostatistics
# TABLE OF CONTENTS

<table>
<thead>
<tr>
<th>Chapter</th>
<th>Title</th>
<th>Page</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Abstract</td>
<td>ii</td>
</tr>
<tr>
<td></td>
<td>Acknowledgments</td>
<td>iii</td>
</tr>
<tr>
<td></td>
<td>Vita</td>
<td>iv</td>
</tr>
<tr>
<td></td>
<td>List of Figures</td>
<td>vii</td>
</tr>
<tr>
<td></td>
<td>List of Tables</td>
<td>ix</td>
</tr>
<tr>
<td></td>
<td>Chapters:</td>
<td></td>
</tr>
<tr>
<td>1.</td>
<td>Introduction</td>
<td>1</td>
</tr>
<tr>
<td>2.</td>
<td>Review of Nucleotide Substitution</td>
<td>8</td>
</tr>
<tr>
<td></td>
<td>2.1 Models of Molecular Evolution</td>
<td>8</td>
</tr>
<tr>
<td></td>
<td>2.2 Synonymous and Nonsynonymous Substitution</td>
<td>15</td>
</tr>
<tr>
<td>3.</td>
<td>Estimating the Synonymous to Nonsynonymous Substitutions Ratio</td>
<td>24</td>
</tr>
<tr>
<td></td>
<td>3.1 The Methodology</td>
<td>25</td>
</tr>
<tr>
<td></td>
<td>3.1.1 Results based on the one-parameter model</td>
<td>27</td>
</tr>
<tr>
<td></td>
<td>3.1.2 Results based on the two-parameter model</td>
<td>37</td>
</tr>
<tr>
<td></td>
<td>3.1.3 Modifications</td>
<td>41</td>
</tr>
<tr>
<td></td>
<td>3.2 Simulation Studies</td>
<td>42</td>
</tr>
<tr>
<td></td>
<td>3.3 Applications</td>
<td>53</td>
</tr>
</tbody>
</table>
## LIST OF FIGURES

<table>
<thead>
<tr>
<th>Figure</th>
<th>Page</th>
</tr>
</thead>
<tbody>
<tr>
<td>1.1</td>
<td>6</td>
</tr>
<tr>
<td>1.1</td>
<td>6</td>
</tr>
<tr>
<td>2.1</td>
<td>10</td>
</tr>
<tr>
<td>2.2</td>
<td>13</td>
</tr>
<tr>
<td>4.1</td>
<td>62</td>
</tr>
<tr>
<td>4.2</td>
<td>69</td>
</tr>
<tr>
<td>4.3</td>
<td>74</td>
</tr>
<tr>
<td>5.1</td>
<td>100</td>
</tr>
<tr>
<td>5.2</td>
<td>101</td>
</tr>
<tr>
<td>5.3</td>
<td>104</td>
</tr>
<tr>
<td>5.4</td>
<td>105</td>
</tr>
</tbody>
</table>
5.5 Topology used to generate sequences for $N = 16$ under null hypothesis

5.6 Results of 100 simulations of statistics $Z_1(5.26)$ and $Z_2(5.27)$ for $N = 16$, the histograms on the first row are based on $m = n = 100$; the second row are based on $m = 100$ and $n = 500$; and the third row are based on $m = n = 500$. .................................................. 107

5.7 Topology used to generate sequences for $N = 32$ under null hypothesis

5.8 Results of 100 simulations of statistics $Z_1(5.26)$ and $Z_2(5.27)$ for $N = 32$, the histograms on the first row are based on $m = n = 100$; the second row are based on $m = 100$ and $n = 500$; and the third row are based on $m = n = 500$. .................................................. 109

5.9 Trees used to generate sequences for $N = 6$. .............................................. 111

5.10 Estimated power curves for $N = 6$. .................................................. 112

5.11 Estimated power curves for $N = 16$ and $32$. .............................................. 114

5.12 The mosaic genome of MAL of HIV-1. The left tree is constructed from the $gag$ gene, and the right tree is from the $env$ gene. ............... 117
LIST OF TABLES

<table>
<thead>
<tr>
<th>Table</th>
<th>Page</th>
</tr>
</thead>
<tbody>
<tr>
<td>1.1</td>
<td>The universal genetic code</td>
</tr>
<tr>
<td>3.1</td>
<td>Primary Parameters and their Estimators in Two-Parameter Model</td>
</tr>
<tr>
<td>3.2</td>
<td>True values of the simulation study</td>
</tr>
<tr>
<td>3.3</td>
<td>Mean and standard deviations of $\hat{K}_S$, $\hat{K}_A$, and $\hat{K}_S/\hat{K}_A$ obtained when the ratio is 1 by excluding inapplicable cases. $n=$ number of applicable cases</td>
</tr>
<tr>
<td>3.4</td>
<td>Mean and standard deviations of $\hat{K}_S$, $\hat{K}_A$, and $\hat{K}_S/\hat{K}_A$ obtained when the ratio is 2 by excluding inapplicable cases. $n=$ number of applicable cases</td>
</tr>
<tr>
<td>3.5</td>
<td>Mean and standard deviations of $\hat{K}_S$, $\hat{K}_A$, and $\hat{K}_S/\hat{K}_A$ obtained when the ratio is 5 by excluding inapplicable cases. $n=$ number of applicable cases</td>
</tr>
<tr>
<td>3.6</td>
<td>Percentage of confidence intervals containing true values when ratio is 1. The confidence intervals for the log($K_S/K_A$) value are the same as the suggested intervals using log transform for the original parameter $K_S/K_A$.</td>
</tr>
<tr>
<td>3.7</td>
<td>Percentages of confidence intervals containing true values when ratio is 2. The confidence intervals for the log($K_S/K_A$) value are the same as the suggested intervals using log transform for the original parameter $K_S/K_A$.</td>
</tr>
</tbody>
</table>
3.8 Percentages of confidence intervals containing true values when ratio is 5. The confidence intervals for the log\((K_S/K_A)\) value are the same as the suggested intervals using log transform for the original parameter \(K_S/K_A\). ..................................................................................................................... 53

3.9 Comparisons of \(D.pseudoobscura\) and \(D.subobscura\). The suggested confidence intervals for the \(K_S/K_A\) ratio are based on the log transform. 55

3.10 \(K_S/K_A\) ratio for \(gag\), the estimated ratio, the SD of the estimator, the suggested 95% confidence intervals for the parameter using the log transform, the mean ratios obtained by MEGA is 5.47. ............................ 57

3.11 \(K_S/K_A\) ratio for \(pol\), the estimated ratio, the SD of the estimator, the suggested 95% confidence intervals for the parameter using the log transform, the mean ratios obtained by MEGA is 7.28. ............................ 58

3.12 Standardized difference between \(gag\) and \(pol\) genes. * indicates the difference is significant at \(\alpha = 0.05\) level. ................................. 59

5.1 Possible numbers of rooted and unrooted trees for 2-10 OTUs....... 84

5.2 Values of \(N, m,\) and \(n\) in the simulation study ............................. 103
CHAPTER 1

Introduction

As more and more deoxyribonucleic acid (DNA) sequences are available, it is inevitable that they will be used in studies of molecular evolution. This dissertation compares the process of evolution between two sets of aligned DNA sequences. The comparisons are made in the categories of nucleotide substitution and phylogenetic inference.

As pointed out by Li and Graur ([37]), molecular evolution encompasses two cases of study: (1) the rates and patterns of changes occurring in the genetic material (e.g., DNA sequences) and its products (e.g., proteins) during evolutionary time and the mechanisms responsible for such changes; (2) the identification of the evolutionary history of genes and organisms. The rapid accumulation of DNA sequence data since the late 1970's has already had a great impact on the study of molecular evolution. They have been used, on the one hand, to estimate the rate of nucleotide substitution, and on the other hand, to infer the phylogenetic relationships among organisms. In this dissertation, we shall explore some interesting questions in both areas.

The hereditary information of all living organisms, with the exception of some viruses, is carried by DNA molecules. DNA usually consists of two complementary chains twisted around each other to form a right-handed helix. Each chain is a linear
polynucleotide consisting of four types of nucleotides. They are the two purines: adenine (A) and guanine (G), and the two pyrimidines: cytosine (C) and thymine (T). A basic process in the evolution of DNA sequences is the change in nucleotide with time. Changes in nucleotide sequences generally take a very long time to complete. For this reason, we cannot deal with the process of nucleotide change by direct observation, and nucleotide changes are typically inferred from pairwise comparisons of DNA molecules that share a common origin. In molecular evolutionary studies, both for estimating the rate of evolution and for reconstructing the evolutionary history of organisms, the changes at different sites of the nucleotide sequences are usually assumed to be independent probabilistic events. This is a restrictive assumption, but practical computation does not appear feasible without it.

DNA sequences are normally copied exactly during the process of chromosome replication. Rarely, however, errors occur that give rise to new sequences. These errors are called mutations. Mutations may be classified by the type of change caused by the mutational event into (1) substitution, the replacement of one nucleotide by another, (2) deletion, the removal of one or more nucleotides from the DNA, (3) insertion, the addition of one or more nucleotides to the sequence, and (4) inversion, the reversal of polarity of a sequence involving two or more nucleotides. The synthesis of proteins involves a process of decoding the nucleotide sequence, whereby the genetic information is translated into amino acids. This translation involves the sequential recognition of codons, which are defined as the adjacent nonoverlapping triplets of nucleotides. Each codon will be translated into a specific amino acid. The correspondence between the codons and the amino acids is determined by a set of rules called the genetic code as shown in Table 1.1. Most amino acids are encoded by more
than one codon. In other words, different codons may be translated into the identical amino acid. Nucleotide substitutions can be characterized by their effect on the product of translation. A substitution is synonymous if it does not cause an amino acid change. Otherwise, it is nonsynonymous. Each of the codons can mutate to nine other codons by means of a single nucleotide substitution. For example, CCT (Pro) can experience six nonsynonymous substitutions, to TCT (Ser), ACT (Thr), GCT (Ala), CTT (Leu), CAT (His), or CGT (Arg), and three synonymous substitutions, to CCC (Pro), CCA (Pro), or CCG (Pro).

When the maximum divergence between taxa is low, distance based on syn-

<table>
<thead>
<tr>
<th>Codon</th>
<th>Amino acid</th>
<th>Codon</th>
<th>Amino acid</th>
<th>Codon</th>
<th>Amino acid</th>
<th>Codon</th>
<th>Amino acid</th>
</tr>
</thead>
<tbody>
<tr>
<td>TTT</td>
<td>Phe</td>
<td>TCT</td>
<td>Ser</td>
<td>TAT</td>
<td>Tyr</td>
<td>TGT</td>
<td>Cys</td>
</tr>
<tr>
<td>TTC</td>
<td>Phe</td>
<td>TCC</td>
<td>Ser</td>
<td>TAC</td>
<td>Tyr</td>
<td>TGC</td>
<td>Cys</td>
</tr>
<tr>
<td>TTA</td>
<td>Leu</td>
<td>TCA</td>
<td>Ser</td>
<td>TAA</td>
<td>Stop</td>
<td>TGA</td>
<td>Stop</td>
</tr>
<tr>
<td>TTG</td>
<td>Leu</td>
<td>TCG</td>
<td>Ser</td>
<td>TAG</td>
<td>Stop</td>
<td>TGG</td>
<td>Trp</td>
</tr>
<tr>
<td>CTT</td>
<td>Leu</td>
<td>CCT</td>
<td>Pro</td>
<td>CAT</td>
<td>His</td>
<td>CGT</td>
<td>Arg</td>
</tr>
<tr>
<td>CTC</td>
<td>Leu</td>
<td>CCC</td>
<td>Pro</td>
<td>CAC</td>
<td>His</td>
<td>CGC</td>
<td>Arg</td>
</tr>
<tr>
<td>CTA</td>
<td>Leu</td>
<td>CCA</td>
<td>Pro</td>
<td>CAA</td>
<td>Gln</td>
<td>CGA</td>
<td>Arg</td>
</tr>
<tr>
<td>CTG</td>
<td>Leu</td>
<td>CCG</td>
<td>Pro</td>
<td>CAG</td>
<td>Gln</td>
<td>CGG</td>
<td>Arg</td>
</tr>
<tr>
<td>ATT</td>
<td>Ile</td>
<td>ACT</td>
<td>Thr</td>
<td>AAT</td>
<td>Asn</td>
<td>AGT</td>
<td>Ser</td>
</tr>
<tr>
<td>ATC</td>
<td>Ile</td>
<td>ACC</td>
<td>Thr</td>
<td>AAC</td>
<td>Asn</td>
<td>AGC</td>
<td>Ser</td>
</tr>
<tr>
<td>ATA</td>
<td>Ile</td>
<td>ACA</td>
<td>Thr</td>
<td>AAA</td>
<td>Lys</td>
<td>AGA</td>
<td>Arg</td>
</tr>
<tr>
<td>ATG</td>
<td>Met</td>
<td>ACG</td>
<td>Thr</td>
<td>AAG</td>
<td>Lys</td>
<td>AGG</td>
<td>Arg</td>
</tr>
<tr>
<td>GTT</td>
<td>Val</td>
<td>GCT</td>
<td>Ala</td>
<td>GAT</td>
<td>Asp</td>
<td>GGT</td>
<td>Gly</td>
</tr>
<tr>
<td>GTC</td>
<td>Val</td>
<td>GCC</td>
<td>Ala</td>
<td>GAC</td>
<td>Asp</td>
<td>GGC</td>
<td>Gly</td>
</tr>
<tr>
<td>GTA</td>
<td>Val</td>
<td>GCA</td>
<td>Ala</td>
<td>GAA</td>
<td>Glu</td>
<td>GGA</td>
<td>Gly</td>
</tr>
<tr>
<td>GTG</td>
<td>Val</td>
<td>GCG</td>
<td>Ala</td>
<td>GAG</td>
<td>Glu</td>
<td>GGG</td>
<td>Gly</td>
</tr>
</tbody>
</table>

Table 1.1: The universal genetic code.
onymous changes may reduce the effect of among-site rate variation, as synonymous substitutions are known to be largely neutral. For more distantly related taxa, restriction to nonsynonymous changes tends to minimize the impact of noise contributed by a large number of silent changes. Hence in the study of the evolutionary divergence of DNA sequences, it is often interesting to estimate the number of synonymous and nonsynonymous nucleotide substitutions separately. This approach of treating synonymous and nonsynonymous substitutions differently is also desirable when the amino acid coding regions of genes are compared, for it is known that the rate of synonymous substitution is generally much greater than that of nonsynonymous substitution.

In this dissertation, we study the rates of synonymous and nonsynonymous substitutions, and the ratio of these rates. The ratio of synonymous mutations per synonymous site to nonsynonymous mutations per nonsynonymous site is a measure of biological functionality that is commonly used by molecular biologists. We propose a method to estimate these three quantities as well as their corresponding variances so that one can compare them among genes or different regions of the same gene, which will help us understand the mechanism of nucleotide substitution in evolution. The model being used allows us to take into account the possibility of having synonymous and nonsynonymous substitutions occurring at the same codon.

Representing evolutionary relationships among a group of organisms by phylogenetic trees is a very important tool in phylogenetic studies. The objects of phylogenetic studies are: (1) to reconstruct the correct genealogical ties between organisms; and (2) to estimate the time of divergence between organisms since they last shared a common ancestor. A phylogenetic tree is a graph composed of nodes and branches,
in which only one branch connects any two adjacent nodes. The nodes represent the
taxonomic units, and the branches define the relationships among the units in terms
of descent and ancestry. The branching pattern of a tree is called the topology. The
branch length usually represents time or the number of changes that have occurred in
that branch. The taxonomic units represented by the nodes can be species, popula-
tions, individuals, or genes.

A node is said to be bifurcating if it has only two immediate descendant lineages,
but multifurcating if it has more than two immediate descendant lineages. Usually,
a phylogenetic tree is assumed to be bifurcating as the simultaneous divergence into
multiple species is thought to be rare. When dealing with phylogenetic trees, we
distinguish between external nodes and internal nodes. External nodes represent the
extant taxonomic units under study and are referred to as operational taxonomic
units (OTUs). Internal nodes represent ancestral units. A tree is said to have a
molecular clock if the rate of molecular evolution is constant over time in all lineages.
Although the rate-constancy assumption has always been controversial, it has been
widely used in the estimation of divergence times and in the reconstruction of phylo-
genetic trees.

Phylogenetic trees can be either rooted or unrooted as illustrated in figure (1.1).
In a rooted tree there exists a particular node called the root, from which a unique
path leads to any other node. The direction of each path corresponds to evolutionary
time, and the root is the common ancestor of all the OTUs under study. An unrooted
tree is a tree that only specifies the relationships among the OTUs and does not define
the evolutionary path.
A phylogenetic tree that represents the evolutionary pathways of a group of species is called a species tree. When a phylogenetic tree is constructed from one gene from each species, the inferred tree is called a gene tree. The gene tree may not always reflect the true phylogenetic relationships among the group of species under study. Numerical methods for inferring phylogenies from molecular data exist. The three major families of methods of inferring phylogeny are the parsimony methods, the distance matrix methods, and the maximum likelihood methods. Most methods fit under one of these headings. The principle of maximum parsimony involves the identification of a tree that requires the smallest number of evolutionary changes to explain the differences observed among the OTUs under study. Often more than one minimizing tree is found, so that no unique tree can be inferred by this method. In the distance matrix methods, evolutionary distances are computed for all pairs of taxa, and a phylogenetic tree is constructed by using an algorithm based on some functional relationships among the distance values. The maximum likelihood estimate of the true tree is the tree that maximizes the probability of obtaining the data given the possible evolutionary tree under a probabilistic model of evolutionary change.
This dissertation will also investigate the issue of whether the phylogenetic relationship of the underlying population remains identical across different sources. One would expect that evolutionary trees for different regions of the genome would be similar, at least within the power of resolution of the analysis. However, this is not always the case. For example, Carr et al. ([2]) reports a full genomic HIV (Human Immunodeficiency Virus) sequence isolated from a patient from Thailand that appears to be a mosaic virus. Sequential analysis along the sequence shows that at first this particular virus appears to derive from the Human Immunodeficiency Virus Type I (HIV-1) subtype A, then the genotype switches to E and back to A for the remaining region. This seeming misclassification may be caused by many reasons. One possibility is recombination, as retroviruses are found to be highly recombinogenic (Hu and Temin [26]). We will develop a general method to discriminate the relationships between two parts of the genome, with which one can detect recombination as a special case.
CHAPTER 2

Review of Nucleotide Substitution

In order to characterize the evolution of a DNA sequence, we need to know how fast it evolves and what are the rates of nucleotide substitution of its constituent parts. One of the three types of genes being recognized is protein-coding genes, which are transcribed into ribonucleic acid (RNA) and subsequently translated into proteins. In studying the evolution of protein-coding genes, it is useful to distinguish between synonymous and nonsynonymous substitutions. However, as the process of nucleotide substitution is usually extremely slow, to detect evolutionary changes in a DNA sequence it is essential to employ comparative methods whereby a given sequence is compared with another sequence with which it shares a common ancestry in the evolutionary past. Such comparisons require statistical methods, some of which will be discussed in this chapter.

2.1 Models of Molecular Evolution

The number of nucleotide substitutions that have occurred since two sequences diverged from each other is one of the most commonly examined variables in the study of molecular evolution. If the degree of divergence is substantial, then the observed number of differences is likely to be smaller than the actual number of substitutions
due to multiple substitutions at the same site. Several methods have been proposed to correct for this distortion and some will be reviewed below. The number of substitutions is usually expressed in terms of the number of substitutions per nucleotide site rather than as the total number of substitutions between the two sequences. This facilitates comparison of the degrees of divergence among sequence pairs that differ in length from each other.

To study the dynamics of nucleotide substitution, we must make several assumptions regarding the probability of substitution of one nucleotide by another. Numerous such mathematical schemes have been proposed in the literature. Most of these models are particular cases of Rodriguez et al.'s ([48]) general 4 hypothesis model (G4H). The G4H model assumes that nucleotide substitution is a Markov process in which the rates of substitution are: (a) not dependent on the sequence site; (b) constant in time; (c) the same for the two lineages; and in which (d) expected base frequencies of the ancestral sequence are equilibrium frequencies, so that they remain unaltered during the process. These four hypotheses are distinct components of the general model. Many other models are particular cases of the G4H model obtained by introducing some additional constraints which simplify the G4H model. We shall restrict our discussion to only the simplest and most frequently used ones: Jukes and Cantor's ([29]) one-parameter model and Kimura's ([32]) two-parameter model.

The Jukes and Cantor ([29]) model assumes that substitutions occur randomly among the four types of nucleotides (Figure 2.1). For example, if the nucleotide under consideration is A, it will change to T, C, or G with equal probability. In this model, the rate of substitution for each nucleotide is $3\alpha$ per unit time, and the rate of substitution in each of the three possible directions of change is $\alpha$. Then what is the
probability that one site is occupied by a particular nucleotide, say A, at any time \( t + 1 \)? Let us first consider this as a discrete-time process and denote this probability to be \( P_{A(t+1)} \).

There are two possible scenarios. (1) The nucleotide has remained unchanged if it is A at time \( t \), and (2) The nucleotide has changed to A if it is not A at time \( t \). The probability of the nucleotide being A at time \( t \) is \( P_A(t) \), and the probability that it has remained A at time \( t + 1 \) is \( 1 - 3\alpha \). The product of these two probabilities gives us the probability for the first scenario. The probability of the nucleotide not being A at time \( t \) is \( 1 - P_A(t) \) and the probability of changing to A at time \( t + 1 \) is \( \alpha \). The product of these two probabilities gives us the probability for the second scenario. Therefore, we have shown the following recurrence equation:

\[
P_{A(t+1)} = (1 - 3\alpha)P_A(t) + \alpha[1 - P_A(t)].
\]

It can be rewritten in terms of the amount of change in \( P_A(t) \) per unit time as

\[
\Delta P_A(t) = P_{A(t+1)} - P_A(t) = -4\alpha P_A(t) + \alpha.
\]
If we approximate this process by a continuous-time model by regarding $\Delta P_{A(t)}$ as the rate of change at time $t$, then the above equation can be rewritten as

$$\frac{dP_{A(t)}}{dt} = -4\alpha P_{A(t)} + \alpha.$$ 

Solving this linear differential equation yields

$$P_{A(t)} = \frac{1}{4} + (P_{A(0)} - \frac{1}{4})e^{-4\alpha t}$$ (2.1)

where $P_{A(0)}$ summarizes the initial conditions. For example, if the initial nucleotide is A, then $P_{A(0)} = 1$, otherwise, $P_{A(0)} = 0$.

We will rewrite the notation $P_{A(t)}$ in a more explicit form to take into account the initial conditions. For example, if the initial nucleotide is C, then the probability that the nucleotide is A at time $t$ will be written as $P_{C||A(t)}$. Since all the nucleotides are equivalent under the Jukes Cantor model, we can consider a more general notation, $P_{i||j(t)}$, which is the probability that a nucleotide has changed to $j$ from initial nucleotide $i$ at time $t$. By using this generalization notation, we obtain

$$P_{i||i(t)} = \frac{1}{4} + \frac{3}{4}e^{-4\alpha t}$$ (2.2)

and

$$P_{i||j(t)} = \frac{1}{4} - \frac{1}{4}e^{-4\alpha t}$$ (2.3)

where $i \neq j$.

This result can be applied to study the nucleotide divergence between two sequences that share a common origin. Let $I(t)$ be the probability that the nucleotide at a given site at time $t$ is the same in both sequences. Suppose the initial nucleotide for both sequences is $i$. Then at time $t$, the probability that a descendant sequence
will remain \( i \) at this site is \( P_{ii(t)} \), and consequentially the probability that both descendant sequences have \( i \) at this site is \( P_{ii(t)}^2 \). Similarly, the probability that both sequences have changed to \( j \) at this site is \( P_{ij(t)}^2 \), where \( j \) could be one of the other three nucleotides. Therefore,

\[
I(t) = P_{ii(t)}^2 + 3P_{ij(t)}^2.
\]  

From equations (2.2) and (2.3), we have

\[
I(t) = \frac{1}{4} + \frac{3}{4}e^{-8\alpha t}.
\]  

Note that the probability that the two sequences are different at a site at time \( t \) is

\[
P = 1 - I(t).
\]

Thus

\[
P = \frac{3}{4}(1 - e^{-8\alpha t})
\]

or equivalently,

\[
8\alpha t = -\ln(1 - \frac{4}{3}P).
\]  

The time of divergence between two sequences is usually not known, and thus we can not estimate \( \alpha \). Under the one-parameter model, the expected number of substitutions per site since the time of divergence between the two sequences, denoted by \( K^* \), is calculated as \( K^* = 2(3\alpha t) \), where \( 3\alpha t \) is the number of substitutions per site in each of the two lineages. By using equation (2.6) we obtain

\[
K^* = -\frac{3}{4}\ln(1 - \frac{4}{3}P).
\]  

where \( P \) is the probability that the two sequences are different at a site. While both \( K^* \) and \( P \) are generally unknown, they can be easily estimated since \( P \) can be estimated by the proportion of different nucleotides between the two sequences under
study. One implicit assumption required so that equation (2.7) can be used is that the proportion of differences between the two sequences can not exceed three quarters.

Of course, the assumption that all nucleotide substitutions occur equally likely is unrealistic in most cases. For example, transitions (substitutions between C and T (pyrimidines) or between A and G (purines) ) are generally more frequent than transversions (substitutions between T,C (a pyrimidine) and A,G (a purine) ). To take this into account, Kimura ([32]) has proposed a two-parameter model (Figure 2.2). In this scheme, the rate of transitional substitution at each nucleotide site is $\alpha$

![Figure 2.2: Two-parameter model](image)

per unit time, whereas the rate of each of the two types of transversional substitution is $\beta$ per unit time. Since transitional substitution is more likely to occur than transversional substitution, $\alpha$ usually has a higher value than $\beta$. The differences between two sequences are classified into transitions and transversions. Let $P$ and $Q$ be the probabilities of transitional and transversional differences at a site between the two sequences, respectively. They can be estimated by their corresponding sample values. Then under the two-parameter model, the expected number of nucleotide
substitutions per site between the two sequences, $K'$, is expressed as

$$K' = -\frac{1}{2} \ln(1 - 2P - Q) - \frac{1}{4} \ln(1 - 2Q). \quad (2.8)$$

As with the one-parameter model case, there is also some restriction on the parameters $P$ and $Q$ to apply the equation (2.8). Generally speaking, both the transitional and transversional differences between the two sequences can not be too large. In the special case of $\alpha = \beta$, i.e., when the rate of transitional substitutions is equal to the rate of each of the two transversional substitutions, the probability of transitional difference $P$ is half of the probability of transversional difference $Q$. Then, substituting $P = Q/2$ in (2.8), we get

$$K' = -\frac{3}{4} \ln(1 - \frac{4}{3} \lambda) \quad (2.9)$$

where $\lambda = P + Q = 3Q/2$ is the probability of any nucleotide difference at one site between the two sequences compared. The result under the two-parameter model becomes the same as that under the simpler one-parameter model. However, in actual situations, $P$ is often larger than $Q$, and therefore the assumption of $\alpha = \beta$ or $P = Q/2$ is not always realistic. When the degree of divergence is small, the two models give essentially the same estimate of $K$. When the degree of divergence is large, however, the estimates by the two models may differ considerably. When this situation arises, and especially in cases where there are a priori reasons to believe that the rate of transition greatly differs from the rate of transversion, the two-parameter model usually tends to be more accurate than the one-parameter model. Other models have also been proposed by Gojobori et al. (15), Tajima and Nei (59), Lanave et al. (34), and Hasegawa et al. (20).
Felsenstein [10] also proposed a simple Markov process model of base substitution to specify \( P_{ij}(t) \), the probability of transition from one base \( i \) to another base \( j \) over a segment of length \( t \).

\[
P_{ij}(t) = e^{-ut} \delta_{ij} + (1 - e^{-ut}) \pi_j
\]

where the stationary probability \( \pi_j \) is the probability that the state is \( j \) after a length of time \( t \), \( u \) is the rate of base substitution per unit time. In other words, the probability of a base changing may depend on its current identity, but not on its past history. The advantage of this method is that the probability of change from base \( i \) to base \( j \) depends on \( t \) only through their product \( ut \), the total substitutions over the length of time \( t \). It was regarded as time measured on a molecular clock which may run at different rates in different segments of the tree. Thus, the total length of segments from the bottom fork up to each tip were not required to be the same.

2.2 Synonymous and Nonsynonymous Substitution

One of the most valuable principles in molecular evolution is that functionally less important proteins or parts of proteins evolve, in terms of substitution rates, faster than the more important ones. In other words, genes or parts of genes that are conserved during the evolutionary process are assumed \textit{a priori} to be functionally more important than those that evolve faster. Because of the large difference between the synonymous and nonsynonymous substitution rates in a gene, in the study of the evolutionary divergence of DNA sequences, it is often important to estimate the synonymous and nonsynonymous nucleotide substitutions rates separately.

Mutations that result in an amino acid substitution have a higher chance of causing deleterious effects on the function of the protein than do synonymous changes.
Consequently, the majority of nonsynonymous mutations will be eliminated from the population by purifying selection. The result will be a reduction in the rate of substitution at nonsynonymous sites. In contrast, synonymous changes have a better chance of being neutral, and more of them will be fixed in a population. Therefore, the rate of synonymous substitution is generally much higher than that of nonsynonymous substitution and is similar for many different genes. Synonymous substitutions may be used as a molecular clock for dating the evolutionary time of closely related species (Kimura [31]; Miyata et al. [43]; Smith et al [57]; Kasper et al [30]). Furthermore, evidence of the selective forces acting on a gene as two alleles diverge can be gathered by estimating the number of synonymous ($K_S$) and nonsynonymous ($K_A$) changes per site. The ratio of $K_S$ to $K_A$ has been used greatly by biologists (Brown and Monaghan [1]; Hughes and Nei [27]; Smith et al [57]; Kasper et al [30]). The ratio of $K_S$ to $K_A$ is of interest in two types of comparisons: (1) $K_S$ is compared to $K_A$ within the same region of DNA sequences, (2) the ratio is compared between different genes or different regions of the same gene. A powerful method of discriminating between positive selection and neutral polymorphism is to compare the rates of synonymous and nonsynonymous nucleotide substitutions per site. If synonymous and nonsynonymous substitutions are selectively equal, then the ratio $K_S/K_A$ is expected to be close to unity; but if purifying selection is preferentially eliminating amino acid changes, then the $K_S/K_A$ ratio will be significantly higher. A higher rate of nonsynonymous substitutions compared with synonymous substitutions ($K_S/K_A < 1$) is evidence of positive selection for amino acid change.

It is a natural extension to employ formulas (2.7) or (2.8) to estimate the number of synonymous substitutions and nonsynonymous substitutions per site. To do so, we
need to replace P and Q with some appropriate probabilities. Regardless of which model is used, we always face two tasks: counting the number of sites for the estimation and the number of differences at each type of site between the two sequences compared.

Generally, each nucleotide site is first classified into a 3/3S (synonymous), 2/3S (2/3 synonymous), 1/3S (1/3 synonymous), or 3/3A (nonsynonymous) site according to the possible synonymous change at that site. For example, the first two positions of the codon TTT (Phe) are counted as nonsynonymous because no synonymous change can occur at these two positions and the third position is counted as one-third synonymous, or identically, as two-third nonsynonymous because one of the three possible changes at this position is synonymous. As for another example, the codon ACT (Thr) has two nonsynonymous sites (the first two positions) and one synonymous site (the third position) because all possible changes at the first two positions are nonsynonymous, and all possible changes at the third position are synonymous.

An alternative way of treating coding regions is to classify nucleotide sites into nondegenerate, twofold degenerate, and fourfold degenerate sites (Li et al. [39]). A site is nondegenerate if all possible changes at this site are nonsynonymous, twofold degenerate if one of the three possible changes is synonymous, and fourfold degenerate if all possible changes at the site are synonymous. For example, the first two positions of the codon TTT (Phe) are nondegenerate while the third position is twofold degenerate (Table 1.1). In comparison the third position of the codon GTT (Val) is fourfold degenerate. The third position in the three isoleucine (Ile) codons is usually treated for simplicity as a twofold degenerate site although in reality the degeneracy at this position is threefold.
Either way, we can count the number of synonymous and nonsynonymous sites easily. Let us now denote the average number of synonymous sites and the average number of nonsynonymous sites between the two sequences compared by $L_S$ and $L_A$, respectively. Later we will explain in detail how they will be obtained in our approach.

Among the many methods that have been proposed to estimate the synonymous and nonsynonymous substitutions, the most frequently used ones are: Li et al.’s (LWL) method ([39]), and Nei and Gojobori’s (NG) method ([44]). These methods agree in general principle while disagreeing in how the weight should be given to two or more possible evolutionary pathways between a pair of codons when the number of differences is counted. Both methods are based on the pattern codon degeneracy in the codon table (Table 1.1). When two sequences are compared, the convention in estimating $K_S$ and $K_A$ is to compare them codon by codon.

In the case of the one-parameter model, if there is only one nucleotide difference between the two codons, it can be immediately decided whether it is synonymous or nonsynonymous. When two nucleotide differences exit, we may then face the following problem. For example, for the two codons AAT and ACG, which are translated into the amino acids Asn and Thr, respectively, there are two possible pathways requiring only two mutations:

Pathway I: AAT(Asn)$\leftrightarrow$ ACT(Thr)$\leftrightarrow$ ACG(Thr)
Pathway II: AAT(Asn)$\leftrightarrow$ AAG(Lys)$\leftrightarrow$ ACG(Thr)

Some authors assume that pathway I and II occur with equal probability. In that case, then the two nucleotide differences consist of a 0.5 synonymous difference and 1.5 nonsynonymous differences by taking the average between these two pathways. Others argue that since synonymous substitutions occur considerably more often than
nonsynonymous substitutions, it is reasonable to assume that path I, which requires one synonymous and one nonsynonymous change, is more likely than path II, which requires two nonsynonymous changes. For example, if a weight of 0.7 is assigned for pathway I and a weight of 0.3 for pathway II, then the number of synonymous differences between the two codons is estimated to be 0.7 \times 1 + 0.3 \times 0 = 0.7, and the number of nonsynonymous differences is 0.7 \times 1 + 0.3 \times 2 = 1.3. Thus, the weighted and unweighted approaches may give somewhat different results. But Nei and Gojobori ([44]) claimed that both methods will give essentially the same estimates in practice since it is rare that two codons will be different by more than one nucleotide in real data. Using either approach, we can estimate the number of synonymous differences (denoted by \( M_S \)) and the number of nonsynonymous differences (denoted by \( M_A \)) between the two coding sequences. Later we will derive our approach to calculate these two variables.

From the above results, the probability of a synonymous difference per synonymous site \( P_S \) is estimated by \( M_S/L_S \), and the probability of a nonsynonymous difference per nonsynonymous site \( P_A \) is estimated by \( M_A/L_A \). Obvious they do not take into account the effect of multiple substitutions at the same site. Formula (2.7) can be used to make such correction to give the expressions for \( K_S \) and \( K_A \) under the one-parameter model

\[
K_S = -\frac{3}{4} \ln (1 - \frac{4}{3} P_S) \tag{2.11}
\]

\[
K_A = -\frac{3}{4} \ln (1 - \frac{4}{3} P_A). \tag{2.12}
\]

More detailed discussion about the process of the estimation can be found in Li and Graur ([37]).
One of the most commonly used models in the study of nucleotide substitution is the two-parameter model proposed by Kimura ([32]). In this scheme, the rates of transitional substitution and transversional substitution are allowed to differ at each nucleotide site. Generally, when the two-parameter model is employed, the two sequences are compared codon by codon and each difference is classified as either a transition or a transversion. It is pointed out by Li et al. ([39]) that substitutions at nondegenerate sites and fourfold degenerate sites are nonsynonymous and synonymous, respectively. At the twofold degenerate sites, generally speaking, transversional substitutions led to nonsynonymous changes, whereas transitional substitutions led to synonymous changes. When computing \( K_S \) and \( K_A \), each fourfold degenerate site is counted as a synonymous site, each nondegenerate site is counted as a nonsynonymous site, and each twofold degenerate site is counted as a one-third synonymous, or identically, a two-third nonsynonymous site.

To allow for the difference between transitional and transversional rates, \( P_i \) and \( Q_i \) are used to denote the probabilities of transitional and transversional differences at an \( i \)-fold degenerate site, respectively. They are estimated by the number of observed transitional and transversional differences, respectively, at \( i \)-fold degenerate sites divided by \( L_i \) (the average of the total number of \( i \)-fold degenerate sites of the two sequences compared; \( i = 0, 2, \) or 4). When there is more than one nucleotide difference between two codons compared, obviously one will again need to make a decision as to how to distribute the weights on different pathways to complete the counting scheme as discussed in the case of the one-parameter model. Instead of just the synonymous and nonsynonymous differences between the two sequences compared, the transitional and transversional differences are also of interest under the two-parameter
model. Using formulas of Kimura ([32]), the expected number of transitional \((A_i)\) and transversional \((B_i)\) substitutions per \(i\)-fold degenerate site are given by

\[
A_i = -\frac{1}{2}\ln(1 - 2P_i - Q_i) + \frac{1}{4}\ln(1 - 2Q_i)
\]

\[
B_i = -\frac{1}{2}\ln(1 - 2Q_i).
\]

The total expected number of substitutions \((K_i)\) per \(i\)th type site is given by \(K_i = A_i + B_i\). Then \(K_S\) and \(K_A\) are obtained by using the appropriate number of weighted differences to number of weighted sites ratios as following

\[
K_S = \frac{L_2A_2 + L_4K_A}{L_2/3 + L_4} \quad (2.13)
\]

\[
K_A = \frac{L_2B_2 + L_0K_S}{2L_2/3 + L_0}. \quad (2.14)
\]

Both Li et al. (LWL) ([39]) and Nei and Gojobori’s (NG) methods ([44]) can considerably overestimate the \(K_S\) value because transitional mutations tend to occur more often than transversional mutations and because most transitional mutations at twofold degenerate sites are synonymous. When nonsynonymous substitutions are studied these methods may give underestimates of \(K_A\). Modified versions of the LWL method (Li [36], Comeron [5]) have been proposed to improve the estimation. The main difference between Li’s method ([36]) and the previous methods is in the weighting of the different fold degenerate sites. For example, a two-fold degenerate site is counted as \(1/3\) synonymous and \(2/3\) non-synonymous site by the previous methods when computing the total synonymous and nonsynonymous sites, which results in using equations (2.13) and (2.14) to estimate \(K_S\) and \(K_A\), respectively. While Li ([36]) proposes to estimate them by

\[
K_S = \frac{L_2A_2 + L_4A_4}{L_2 + L_4} + B_4 \quad (2.15)
\]
\[ K_A = A_0 + \frac{L_2 B_2 + L_0 B_0}{L_2 + L_0}. \] (2.16)

It is found that the estimation of \( K_S \) by Li's new method is better than that by the old version, while the estimation of \( K_A \) is about the same.

Ina [28] conducted a computer simulation study to evaluate the accuracies of different methods used to estimate \( K_S \) and \( K_A \). The methods being compared include: Nei and Gojobori's (NG) method, Miyata and Yasunaga's (MY) method, Li, Wu, and Luo’s (LWL) method, and Li's ([36]) method. It was found that the NG, MY, and LWL methods give overestimates of \( K_S \) and underestimates of \( K_A \). The major cause of the biased estimation is that these three methods underestimate the number of synonymous sites and overestimate the number of nonsynonymous sites according to the study. The study also concludes that the more recently proposed Li's ([36]) method gives better estimates of the \( K_S \) and \( K_A \) values than those obtained by the NG, MY, and LWL methods. However, when there are strong transition/transversion and nucleotide-frequency biases like mitochondrial genes, all the methods are found to give biased estimations of substitution numbers.

Although many methods have been developed to provide the estimation of \( K_S \) and \( K_A \) and their corresponding variance, there is little effort devoted to assess the variation of the ratio. Consequentially, both comparisons mentioned earlier in the chapter are usually made without statistical justification. When \( K_S \) and \( K_A \) are compared, a two-sample test is often used by mistakenly ignoring the correlation between these two estimators. And due to the lack of the estimate of the variance of the \( K_S/K_A \) ratio, a test of the ratio being equal to a constant is not able to be carried out either. For exactly the same reason, the comparison of the ratio between different genes or different regions of the same gene has not been done convincingly.
We will propose a new way for estimating $K_S$, $K_A$, and their ratio $K_S/K_A$ in next chapter. Estimates of the variance of all three variables will also be provided.
CHAPTER 3

Estimating the Synonymous to Nonsynonymous Substitutions Ratio

In this chapter we propose a new method to estimate the number of synonymous, nonsynonymous substitutions per site, and their ratio. The evaluation of the variation of these estimates is also provided. Results based on both Jukes and Cantor’s ([29]) one-parameter model, and Kimura’s ([32]) two-parameter model are provided. Due to the simplicity of the one-parameter model and the similarity between the two models during the process of developing this method, detailed theoretical consideration has been given to the case of the one-parameter model. The case of the two-parameter model is treated as an extension of the method for the simpler one-parameter model. Computer simulations are conducted to assess the accuracy of our method. We shall also apply this new method to two examples to further demonstrate its usage.

The method is based on some large sample approximations. Each nucleotide site is regarded as a sample unit of the long DNA sequences. In making comparison along the two sequences, we assume that the mutations at different sites are independent events. This assumption is essential in molecular studies. It also provides a valid base for the usage of the central limit theorem in our research.
When estimating the number of mutations between the two sequences compared by either the one-parameter or two-parameter model, the implicit assumption is that the proportion of differences can not be too large. This assumption is undoubtedly inherited by the proposed approach, as well as all other methods that employ these two models. Therefore, when the mutation rate between the two sequences is extremely high, the proposed methods may fail to make an estimation.

3.1 The Methodology

In our approach, we first classify each nucleotide site along the two aligned sequences compared into a 3/3S (synonymous), 2/3S (2/3 synonymous), 1/3S (1/3 synonymous), or 3/3A (nonsynonymous) site according to the possible synonymous changes at that site as described in the previous chapter. The average number of the \( i \)th type of site between the two sequences, denoted by \( N_i \), is then counted. Here \( i = 0 \) for 3/3A, 1 for 1/3S, 2 for 2/3S, or 3 for 3/3S. Hence, \( i \) indexes the possible synonymous changes at a particular site. We then follow different independent processes at each of these four types of sites separately to count the number of differences between the two sequences. The counting scheme will also depend on the model we choose to use.

In the case of the one-parameter model, we first classify the nucleotide difference into the various types of synonymous and nonsynonymous differences. For example, the difference between CAC(His) and CAA(Gln) is labeled as nonsynonymous at a 1/3 synonymous site, while the difference between CAG(Gln) and CAA(Gln) is labeled as synonymous at a 1/3 synonymous site. The purpose of such classification is to enable one to treat synonymous and nonsynonymous substitution at different
types of sites differently. A synonymous substitution at a 1/3 synonymous site and a synonymous substitution at a 2/3 synonymous site will have a different impact on the estimate of the substitution rates and their variances. The average number of synonymous \((M_{i,S})\) or nonsynonymous \((M_{i,A})\) substitutions at the \(i\)th type of site is computed.

In the case of the two-parameter model, we will also classify the nucleotide differences into transitional or transversional differences. For example, the difference between CAC(His) and CAA(Gln) is now labeled as a transversional difference at a 1/3 synonymous site, while the difference between CAG(Gln) and CAA(Gln) is labeled as a transitional difference at a 1/3 synonymous site. Then \(A_i\), the average number of transitional substitutions, and \(B_i\), the average number of transversional substitutions at the \(i\)th type of site will be computed.

We now introduce some notation that will be used frequently in this chapter. Let \(L\) be the length of the sequence, in other words, the number of total sites of each sequence; \(L_S\) be the average number of total synonymous sites, which is equal to \(N_3 + 2/3N_2 + 1/3N_1\); and \(L_A\) be the average number of total nonsynonymous sites, which is equal to \(N_0 + 2/3N_1 + 1/3N_2\). Therefore, \(L_S\) and \(L_A\) can be regarded as the "weighted" total number of synonymous and nonsynonymous sites between the two sequences compared, respectively. For example, a 3/3 synonymous site is assigned a weighted score of 3/3, while a 2/3 synonymous site is assigned a weighted score of 2/3 when counting the synonymous sites. One can easily see that \(L = N_3 + N_2 + N_1 + N_0 = L_S + L_A\).
In our approach, we are going to treat the $N_i$'s ($i = 0, 1, 2, 3$) as fixed to simplify the calculation even though they can take on random values for the sequences of identical length. However, we feel this conditional approach alters the inference very little since the variations of these quantities are relatively much smaller compared to those of the mutations. It is also worthy of pointing out that $M_{2,S}$ and $M_{2,A}$ are counting different random events occurring at the same sites, hence the correlation between these two variables should be taken into account when estimating the variances. This is also the case with $M_{1,S}$ and $M_{1,A}$, $A_1$ and $B_1$, etc.

### 3.1.1 Results based on the one-parameter model

We first consider synonymous substitution under the one-parameter model. Given any two aligned DNA sequences, the probability of synonymous substitutions per site along the sequence can be expressed as

$$p_s = \frac{N_3p_{3,s} + \frac{2}{3}N_2p_{2,s} + \frac{1}{3}N_1p_{1,s}}{L_s}$$

(3.1)

where $p_{i,s}$ is the probability of synonymous substitution at an $i$th type site. When computing the synonymous substitution rate, both the denominator and numerator are weighted. This in fact is crucial when making estimation. Some of the methods currently used only assign different weights to the denominator but not to the numerator, which results in overestimating the value of $p_s$, and hence that of $K_s$. Because of the independence between different sites, the number of synonymous differences at an $i$th type of site, $M_{i,S}$, follows a binomial distribution with $N_i$ trials and probability of success $p_{i,s}$ for each trial, i.e., $M_{i,S} \sim Bin(N_i, p_{i,s})$. The maximum likelihood estimator for $p_{i,s}$ is $\frac{M_{i,S}}{N_i}$. Here we only consider the cases when $i = 1, 2, 3$, since a 3/3A site, the case when $i = 0$, does not contribute to synonymous mutation according to
the classification of the sites. Continuing to condition on \( N_i \), under some mild conditions, \( M_{i,S} \) is approximately normally distributed with mean \( N_i p_{i,S} \) and variance \( N_i p_{i,S} q_{i,S} \), equivalently we can say

\[
\frac{i}{3} M_{i,S} \sim N\left(\frac{i}{3} N_i p_{i,S}, \frac{i^2}{32} N_i p_{i,S} q_{i,S}\right)
\]

where \( q_{i,S} \) is the conventional notation for \( 1 - p_{i,S} \). These mild conditions are that \( N_i \) should be large and \( p_{i,S} \) should not be extreme (extreme being near 0 or 1). A conservative rule to follow is that the approximation will be good if \( \min(N_i p_{i,S}, N_i (1 - p_{i,S})) \geq 5 \) (Casella and Berger[3]).

Since \( M_{3,S}, M_{2,S}, \) and \( M_{1,S} \) are random events occurring at 3/3S, 2/3S, and 1/3S sites respectively, and these three types of sites are classified exclusively, the random variables \( M_{3,S}, M_{2,S}, \) and \( M_{1,S} \) are indeed independent. Therefore,

\[
\hat{P}_S = \frac{M_{3,S} + \frac{2}{3} M_{2,S} + \frac{1}{3} M_{1,S}}{L_S} \sim N(P_S, \sigma_{P_S}^2)
\]

(3.2)

where \( P_S \) is defined in (3.1), and

\[
\sigma_{P_S}^2 = \frac{N_3 p_{3,S} q_{3,S} + \frac{2^2}{3^2} N_2 p_{2,S} q_{2,S} + \frac{1^2}{3^2} N_1 p_{1,S} q_{1,S}}{L_S^2}.
\]

(3.3)

The probability of synonymous differences per site, \( P_S \), can thus be unbiasedly estimated by \( \hat{P}_S \), the ratio of the sum of weighted synonymous mutations to the sum of weighted synonymous sites. The similar result for nonsynonymous substitutions under the one-parameter model can also be deduced:

\[
\hat{P}_A = \frac{M_{0,A} + \frac{2}{3} M_{1,A} + \frac{1}{3} M_{2,A}}{L_A} \sim N(P_A, \sigma_{P_A}^2)
\]

(3.4)

where

\[
P_A = \frac{N_0 p_{0,A} + \frac{2}{3} N_1 p_{1,A} + \frac{1}{3} N_2 p_{2,A}}{L_A}
\]

(3.5)
and

$$\sigma^2_{P_A} = \frac{N_0 P_0 \cdot A q_{0,A} + \frac{22}{3^2} N_1 P_1 \cdot A q_{1,A} + \frac{12}{3^2} N_2 P_2 \cdot A q_{2,A}}{L_A^2}. \quad (3.6)$$

In our setting, $P_S$ and $P_A$ are the probabilities of synonymous and nonsynonymous differences per site between the two sequences compared, respectively. Their corresponding maximum likelihood estimators are $\hat{P}_S$ and $\hat{P}_A$. Further calculation is based on the well-known invariance property of the maximum likelihood estimator which, informally speaking, says that if $\hat{\theta}$ is the maximum likelihood estimator of $\theta$, then the maximum likelihood estimator of some function of $\theta$, say $\tau(\theta)$, is $\tau(\hat{\theta})$. According to this property, then the maximum likelihood estimators of $K_S = -3/4 \ln(1 - 4/3 P_S)$, and $K_A = -3/4 \ln(1 - 4/3 P_A)$ are

$$\hat{K}_S = -3/4 \ln(1 - \hat{P}_S) \quad (3.7)$$
$$\hat{K}_A = -3/4 \ln(1 - \hat{P}_A). \quad (3.8)$$

It is straightforward now to calculate the approximate variances of $\hat{K}_S$ and $\hat{K}_A$ by applying the delta method, which yields

$$Var(\hat{K}_S) \approx \left( \frac{3}{3 - 4 P_S} \right)^2 \sigma^2_{P_S} \quad (3.9)$$
$$Var(\hat{K}_A) \approx \left( \frac{3}{3 - 4 P_A} \right)^2 \sigma^2_{P_A}. \quad (3.10)$$

To derive the variance of $\hat{K}_S/\hat{K}_A$, we also need to know the covariance between $\hat{P}_S$ and $\hat{P}_A$. It is computed as if $M_{i,S}$ and $M_{i,A}$ are multinomial distributed ($i = 0,1,2,3,4$), i.e., at a particular site, there can be either a synonymous difference, a nonsynonymous difference, or no difference between the two sequences.

$$cov(\hat{P}_S, \hat{P}_A) = \frac{1}{L_S L_A} cov(M_{3,S} + \frac{2}{3} M_{2,S} + \frac{1}{3} M_{1,S}, M_{0,A} + \frac{2}{3} M_{1,A} + \frac{1}{3} M_{2,A})$$

29
\[
\frac{1}{L_S L_A} \left( \text{cov}(\frac{2}{3} M_{2,S}, \frac{1}{3} M_{2,A}) + \text{cov}(\frac{1}{3} M_{1,S}, \frac{2}{3} M_{1,A}) \right) - \frac{2}{9L_S L_A} (N_2 p_{2,S} + N_1 p_{1,S} - N_2 p_{2,A} + N_1 p_{1,A}).
\] (3.11)

The covariance between \( \hat{P}_S \) and \( \hat{P}_A \) can be argued to be negligible since \( 2/3S \) and \( 1/3S \) are usually the minority sites along the sequences. However, we will still take the covariance into account when developing the estimators. First, we will introduce a lemma that will be used later.

**Lemma 3.1.1** Let \( X_n \) be a multivariate random variable with mean \( \mu_n \) and covariance matrix \( \Sigma_n \), which satisfy

\[
\lim_{n \to \infty} \mu_n = \mu
\]
\[
\lim_{n \to \infty} \Sigma_n = \Sigma
\]

In addition, if every nontrivial linear combination of the components of \( X_n \) is asymptotically normally distributed, i.e., \( \forall u \), whose components are not all zeros, if

\[
u' X_n \overset{d}{\to} N(\mu u, \nu' \Sigma u)
\]

then \( X_n \) is asymptotically normally distributed with mean \( \mu \) and covariance matrix \( \Sigma \), namely,

\[
X_n \overset{d}{\to} N(\mu, \Sigma)
\] (3.12)

**Proof:**

Since \( u' X_n \overset{d}{\to} N(u \mu, u' \Sigma u) \), then by theorem (2.2) of Durrett [8], for every bounded continuous function \( g \), we have \( Eg(u' X_n) \to Eg(N(u \mu, u' \Sigma u)) \). We take \( g(x) = e^{ix} \),

30
then the characteristic function of a linear combination $u'X_n$ is

$$Ee^{it u'X_n} \rightarrow e^{it u'\mu - \frac{1}{2}t^2 u'\Sigma u}.$$  

Now setting $t = 1$, we have

$$Ee^{i u'X_n} \rightarrow e^{i u'\mu - \frac{1}{2}u'\Sigma u}.$$  

The right-hand side is the characteristic function of $N(\mu, \Sigma)$, and the result is proved.

□

This lemma is an extension of a well-known property in multivariate analysis. That property states that a multivariate variable is multivariate normal distributed if all linear combinations of its components follow an univariate normal distribution.

Weaker assumptions are made for this lemma, namely, only asymptotic normality is required for the linear combination of the variable's components. In return, the multivariate variable will be asymptotically normal distributed. Now, we will prove the asymptotic normality of the bivariate variable

$$\begin{bmatrix} \hat{p}_{A-P_A} - \hat{p}_A \\ \hat{p}_{S-P_S} - \hat{p}_S \\ \hat{p}_{P_{A}} - \hat{p}_A \\ \hat{p}_{P_{S}} - \hat{p}_S \end{bmatrix}$$

by using lemma (3.1.1).

**Theorem 3.1.2** If

$$\lim_{L \to \infty} \frac{N_i}{L} = \alpha_i > 0$$

for $i = 0, 1, 2, 3$, the probabilities of synonymous and nonsynonymous mutations at all four types of sites are between 0 and 1, and the inequality

$$R$$

$$= \frac{2(\alpha_2 p_{2.2} p_{.A} + \alpha_1 p_{1.1} p_{.A})}{9 \sqrt{\alpha_3 p_{3.3} p_{.S} + \frac{2^2}{3^2} \alpha_2 p_{2.2} p_{.S}} + \frac{1^2}{3^2} \alpha_1 p_{1.1} p_{.S} \sqrt{\alpha_0 p_{0.0} p_{.A} + \frac{2^2}{3^2} \alpha_1 p_{1.1} p_{1.1} + \frac{1^2}{3^2} \alpha_2 p_{2.2} p_{2.2}}}$$

$$< 1$$

(3.13)
is satisfied, then the two-dimension random variable
\[
\begin{bmatrix}
\hat{P}_S - P_S \\
\sigma_{P_S}
\end{bmatrix}
\begin{bmatrix}
\hat{P}_A - P_A \\
\sigma_{P_A}
\end{bmatrix}
\]
is asymptotically bivariate normal distributed with mean \([0, 0]\) and covariance matrix
\[
\begin{bmatrix}
1 & -R \\
-R & 1
\end{bmatrix}
\].

**Proof:** We start with showing that the mean vector and covariance matrix of
\[
\begin{bmatrix}
\hat{P}_S - P_S \\
\sigma_{P_S}
\end{bmatrix}
\begin{bmatrix}
\hat{P}_A - P_A \\
\sigma_{P_A}
\end{bmatrix}
\]
are convergent. From (3.2) and (3.4), it is obvious that
\[
E\left(\frac{\hat{P}_S - P_S}{\sigma_{P_S}}\right) \to 0 \quad (3.14)
\]
\[
E\left(\frac{\hat{P}_A - P_A}{\sigma_{P_A}}\right) \to 0 \quad (3.15)
\]
\[
\text{Var}\left(\frac{\hat{P}_S - P_S}{\sigma_{P_S}}\right) \to 1 \quad (3.16)
\]
\[
\text{Var}\left(\frac{\hat{P}_S - P_S}{\sigma_{P_S}}\right) \to 1 \quad (3.17)
\]
\[
cov\left(\frac{\hat{P}_S - P_S}{\sigma_{P_S}}, \frac{\hat{P}_A - P_A}{\sigma_{P_A}}\right) = \frac{\text{cov}(\hat{P}_S, \hat{P}_A)}{\sigma_{P_S} \sigma_{P_A}}. \quad (3.18)
\]
Following equations (3.3), (3.6), and (3.11), the covariance between \(\frac{\hat{P}_S - P_S}{\sigma_{P_S}}\) and \(\frac{\hat{P}_A - P_A}{\sigma_{P_A}}\) becomes
\[
\frac{2(N_2p_2,s_1,p_2,a + N_1p_1,s_1,p_1,a)}{9\sqrt{N_3p_3,s_3,s_3} + \frac{22}{32}N_2p_2,s_2,s + \frac{12}{32}N_1p_1,s_1,s\sqrt{N_0p_0,a,q_0,a} + \frac{22}{32}N_1p_1,q_1,a + \frac{12}{32}N_2p_2,a,q_2,a}
\]
which is convergent to \(-R\) where \(R\) is defined in (3.13). Therefore, we have shown that the mean vector and covariance matrix of
\[
\begin{bmatrix}
\hat{P}_S - P_S \\
\sigma_{P_S}
\end{bmatrix}
\begin{bmatrix}
\hat{P}_A - P_A \\
\sigma_{P_A}
\end{bmatrix}
\]
is asymptotically bivariate normal distributed.
converge to \[
\begin{bmatrix} 0 \\ 0 \end{bmatrix}
\] and \[
\begin{bmatrix} 1 & -R \\ -R & 1 \end{bmatrix}
\] respectively. Hence to prove the asymptotic bivariate normality, it is now sufficient to show that any linear combination of \( \frac{\hat{p}_2-p_2}{\sigma_{p_2}} \) and \( \frac{\hat{p}_1-p_1}{\sigma_{p_1}} \), \( a\frac{p_2-p_2}{\sigma_{p_2}} + b\frac{p_1-p_1}{\sigma_{p_1}} \), is asymptotically normal distributed. Here we only consider the nontrivial cases when neither \( a \) nor \( b \) is 0.

Let

\[
\hat{S}^j = M_{3/3S,S}^j + \frac{2}{3} M_{2/3S,S}^j + \frac{1}{3} M_{1/3S,S}^j \\
\hat{A}^j = M_{3/3A,A}^j + \frac{2}{3} M_{1/3A,A}^j + \frac{1}{3} M_{2/3A,A}^j
\]

denote the observed weighted synonymous and nonsynonymous differences for the \( j \)th codon, respectively; and

\[
S^j = N_{3/3S}^j p_3/3S.S + \frac{2}{3} N_{2/3S}^j p_2/3S.A + \frac{1}{3} N_{1/3S}^j p_1/3S.A \\
A^j = N_{3/3A}^j p_3/3A.A + \frac{2}{3} N_{1/3A}^j p_1/3A.A + \frac{1}{3} N_{2/3A}^j p_2/3A.A
\]

denote the probabilities of synonymous and nonsynonymous substitution for the \( j \)th codon. Where \( M_{i,S}^j \) and \( N_i^j \) are defined the same as \( M_{i,S} \) and \( N_i \) but restricted to the \( j \)th codon only. The notations \( \hat{S}, \hat{A}, S, \) and \( A \) are used to denote the corresponding sums of these quantities along the sequences. We will first define

\[
\sigma_S^2 \equiv \sigma_{p_2}^2 L_S^2 \\
\quad = N_{3}^j p_3 Aq_3.A + \frac{2}{3} N_{2}^j p_2 Aq_2.A + \frac{1}{3} N_{1}^j p_1 Aq_1.A \tag{3.19}
\]

\[
\sigma_A^2 \equiv \sigma_{p_1}^2 L_A^2 \\
\quad = N_{0}^j p_0 Aq_0.A + \frac{2}{3} N_{1}^j p_1 Aq_1.A + \frac{1}{3} N_{2}^j p_2 Aq_2.A \tag{3.20}
\]
We shall employ the Lindeburg-Feller theorem to justify the asymptotic normality of \( a\hat{S} + b\hat{A} \). Consider

\[
Z_j = \frac{\hat{S}^{ij} - S^i}{\sigma_S} + b\frac{\hat{A}^j - A^j}{\sigma_A}
\]

(3.21)

where \( j = 1, \ldots, L/3 \), is the index of the codon along the two sequences compared. Then by the assumption of independence of the sites, the \( Z_j \)'s are independent random variables with mean 0 since \( \hat{S}^i \) and \( \hat{A}^j \) are unbiased estimators of \( S^i \) and \( A^j \), respectively.

\[
EZ_j^2 = E[(a\frac{\hat{S}^i - S^i}{\sigma_S} + b\frac{\hat{A}^j - A^j}{\sigma_A})^2]
\]

\[
= var(a\frac{\hat{S}^i}{\sigma_S} + b\frac{\hat{A}^j}{\sigma_A})
\]

\[
= a^2\frac{var(\hat{S}^i)}{\sigma_S^2} + b^2\frac{var(\hat{A}^j)}{\sigma_A^2} + 2ab\frac{cov(\hat{S}^i, \hat{A}^j)}{\sigma_S\sigma_A}.
\]

Taking sums of the above equation over all \( j \), we have

\[
\sum_{j=1}^{L/3} EZ_j^2 = a^2 + b^2 + \frac{2ab cov(\hat{S}, \hat{A})}{\sigma_S\sigma_A}
\]

(3.22)

Following equations (3.11), (3.19), and (3.20), the right-hand side of equation (3.22) can be rewritten as

\[
a^2 + b^2 + \frac{4ab(N_2p_2,p_2,A + N_1p_1,p_1,A)}{9\sqrt{N_3p_3,sq_3,s} + \frac{2^2}{3^2}N_2p_2.sq_2.s + \frac{2^2}{3^2}N_1p_1.sq_1.s \sqrt{N_0p_0,aq_0,0,A + \frac{2^2}{3^2}N_1p_1,aq_1,0,A + \frac{12}{3^2}N_2p_2,aq_2,A}}
\]

and it will be convergent to

\[
\sigma^2 \triangleq a^2 + b^2 - 2abR.
\]

(3.23)

Clearly, \( \sigma^2 \) is positive when \( ab < 0 \) since \( R > 0 \) by the definition of \( R \) from (3.13). As for the cases when \( ab > 0 \), without loss of generality, we will assume that both \( a \) and
\( b \) are positive. Then again from the condition (3.13) of the theorem, we know that

\[
\sigma^2 > a^2 + b^2 - 2ab \\
\geq 0. \tag{3.24}
\]

So far, we have shown that for all nontrivial pairs of \((a, b)\), equation (3.23) yields a positive number. Next, we will check the last condition of the Lindeburg-Feller theorem. Here we use the notation \( E(x; A) \) for the expected value of a random variable \( x \) given the statistical event \( A \).

\[
\sum_{j=1}^{L/3} E(|Z_j|^2; |Z_j| > \epsilon) = \frac{L}{3} E(|Z|^2; |Z| > \epsilon)
\]

\[
\leq \frac{L}{3} \left[ E\left(\frac{\hat{S} - S}{\sigma_S}^2; \frac{|a \hat{S} - S|}{\sigma_S} + \frac{b \hat{A} - A}{\sigma_A} > \epsilon\right) \\
+ E\left(\frac{|2ab \frac{\hat{S} - S}{\sigma_S} \hat{A} - A}{\sigma_A} |; \frac{|a \hat{S} - S|}{\sigma_S} + \frac{b \hat{A} - A}{\sigma_A} > \epsilon\right) \\
+ E\left(\frac{b \hat{A} - A}{\sigma_A}^2; \frac{|a \hat{S} - S|}{\sigma_S} + \frac{b \hat{A} - A}{\sigma_A} > \epsilon\right)\right]
\]

\[
\leq \frac{L}{3} \left[ a^2 \sigma_S E\left(|\hat{S} - S|^2; |\hat{S} - S| > \frac{\sigma_S \epsilon}{a} - \frac{b \sigma_S}{a \sigma_A}\right) \\
+ \frac{2ab}{\sigma_S \sigma_A} E\left(\frac{|\hat{S} - S|}{\sigma_S \sigma_A} (\hat{A} - A); |\hat{S} - S| > \frac{\sigma_S \epsilon}{a} - \frac{b \sigma_S}{a \sigma_A}\right) \\
+ \frac{b^2}{\sigma_A^2} E\left(\frac{\hat{A} - A}{\sigma_A}^2; \frac{|\hat{S} - S|}{\sigma_S} + \frac{b \hat{A} - A}{\sigma_A} > \frac{\sigma_A \epsilon}{b} - \frac{a \sigma_A}{b \sigma_S}\right)\right]
\]

By the definition of \( \hat{S}_i, \hat{A}_i, S^i, \) and \( A^i \), it is clear that \( E(|\hat{S} - S|^2) \), \( E(\frac{|\hat{S} - S|}{\sigma_S} (\hat{A} - A)) \), and \( E(\frac{\hat{A} - A}{\sigma_A}^2) \) are all finite. According to the conditions of the theorem, as \( L \rightarrow \infty \), both \( \sigma_S \) and \( \sigma_A \) will approach \( \infty \) as well, while the ratios of
$L/\sigma_S^2$, $L/\sigma_S\sigma_A$, and $L/\sigma_A^2$ are going to 0. Therefore

$$\lim_{L \to \infty} \sum_{j=1}^{L/3} E(|Z_j|^2; |Z_j| > c) \to 0.$$  

(3.25)

Combining (3.22) and (3.25), we conclude that the sum of $Z_j$ is asymptotically normally distributed with mean 0 and variance $\sigma^2$ by the Lindeburg-Feller theorem. For any $a$ and $b$, we notice that

$$\sum Z_j = \sum (a \frac{S_i - S_i^j}{\sigma_S} + b \frac{A_i - A_i^j}{\sigma_A})$$

$$= (a \frac{\hat{S} - S}{\sigma_S} + b \frac{\hat{A} - A}{\sigma_A})$$

$$= \frac{L_S(\hat{S} - S)}{L \sigma_S} + \frac{L_A(\hat{A} - A)}{L \sigma_A}$$

$$= \frac{\hat{P}_S - P_S}{\sigma_{P_S}} + \frac{\hat{P}_A - P_A}{\sigma_{P_A}}.$$  

(3.26)

Because $a$ and $b$ are arbitrary, we draw the conclusion that every nontrivial linear combination of $\frac{\hat{P}_S - P_S}{\sigma_{P_S}}$ and $\frac{\hat{P}_A - P_A}{\sigma_{P_A}}$ is asymptotically normally distributed with mean 0 and variance $\sigma^2$. To complete the proof by applying lemma(3.1.1), we only need to show

$$\sigma^2 = \begin{bmatrix} a & b \end{bmatrix} \begin{bmatrix} 1 & -R \\ -R & 1 \end{bmatrix} \begin{bmatrix} a \\ b \end{bmatrix}$$

which can be easily done following the definition of $\sigma^2$ from (3.23). □

Though sometimes it is difficult to check the validity of inequality (3.13), it is actually a very weak condition that needs to be satisfied. In practice, $\alpha_1$ and $\alpha_2$, the numbers of 1/3 synonymous and 2/3 synonymous sites respectively, are usually much smaller than $\alpha_0$ and $\alpha_3$, the numbers of the other two types of sites. Therefore, inequality (3.13) can be regarded as always being true in practice. Writing out the result of the previous theorem in a slightly different form, we have

$$[\sqrt{L}(\hat{P}_S - P_S), \sqrt{L}(\hat{P}_A - P_A)] \overset{appr.}{\rightarrow} N(0, \Sigma)$$

36
where
\[ \Sigma = \begin{bmatrix} \sigma^2_{PS} & \text{Lcov}(\hat{P}_S, \hat{P}_A) \\ \text{Lcov}(\hat{P}_S, \hat{P}_A) & \sigma^2_{PA} \end{bmatrix}. \]

To derive the final result, we should employ the two dimensional delta-method, which states the following:

**Suppose that**
\[ [\sqrt{n}(X_1 - \theta_1), \sqrt{n}(X_2 - \theta_2)] \to N(0, \Sigma) \]

\( f_1, f_2 \) are real-valued functions of \( \theta = (\theta_1, \theta_2) \), and define the matrix \( B \) of partial derivatives.

Let \( X = (X_1, X_2) \), then
\[ [\sqrt{n}(f_1(X) - f_1(\theta)), \sqrt{n}(f_2(X) - f_2(\theta))] \to N(0, B\Sigma B^\prime) \]

Now we take \( \theta_1, \theta_2, X_1, \) and \( X_2 \) to be \( P_S, P_A, \hat{P}_S, \) and \( \hat{P}_A \), respectively. If we also let \( f_1(K_S, K_A) = \frac{K_x}{K_A} = \frac{\ln(1-P_S)}{\ln(1-P_A)} \), and \( f_2(K_S, K_A) = K_S \), then the variance of \( \hat{K}_S/\hat{K}_A \) can then be shown to be
\[ \text{var}(\frac{\hat{K}_S}{\hat{K}_A}) \approx c_1^2 \sigma^2_{PS} + 2c_1c_2 \text{cov}(\hat{P}_S, \hat{P}_A) + c_2^2 \sigma^2_{PA} \quad (3.27) \]

where
\[ c_1 = -\frac{4}{(3 - 4P_S) \ln(1 - 4/3P_A)} \]
\[ c_2 = \frac{4 \ln(1 - 4/3P_S)}{(3 - 4P_A) \ln(1 - 4/3P_A))^2}. \]

### 3.1.2 Results based on the two-parameter model

The strategy used to develop the results for the one-parameter model is extended in this section. The main difference between the one-parameter model and the two-parameter model is that the former assumes equal rates for all kinds of mutations.
while the later one allows for unequal rates between transitional and transversional substitutions. This difference makes the computation more complicated when applying the two-parameter model since there are more parameters involved in the estimation. When the one-parameter model is employed, the primary parameters of interest are the probabilities of synonymous \( P_s \), defined in (3.1)) and nonsynonymous \( P_A \), defined in (3.5)) substitutions per site. When the two-parameter model is used, four parameters are introduced in order to apply (2.8), namely, \( P_s \) [\( P_A \)] and \( Q_s \) [\( Q_A \)] : the probabilities of transitional and transversional differences, respectively, that lead to synonymous [nonsynonymous] substitutions.

Substitutions at 3/3S sites and 3/3A sites are synonymous and nonsynonymous, respectively. At the 2/3S and 1/3S sites, generally speaking, transitions lead to synonymous changes, while transversions lead to nonsynonymous changes (Li et al. [39]). In other words, synonymous mutation can be caused by either transitions at 1/3S, 2/3S, 3/3S sites, or transversions at 3/3S sites, while nonsynonymous mutation can be caused by the transversions at 1/3S, 2/3S, 3/3A sites, or transitions at 3/3A sites.

This fact is formulated by the following equations:

\[
\begin{align*}
    P_s &= \frac{N_3 p_{3,A} + \frac{2}{3} N_2 p_{2,A} + \frac{1}{3} N_1 p_{1,A}}{L_s} \\
    Q_s &= p_{3,B} \\
    P_A &= p_{0,A} \\
    Q_A &= \frac{N_0 p_{0,B} + \frac{2}{3} N_1 p_{1,B} + \frac{1}{3} N_2 p_{2,B}}{L_A}
\end{align*}
\tag{3.28, 3.29, 3.30, 3.31}
\]

where \( p_{i,A} \) [\( p_{i,B} \)] is the probability of a transitional [transversional] substitution at an \( i \)th type of site.
Analogous to the analysis of the one-parameter model, the estimators of these four parameters can be shown to be approximately normally distributed with means and variances as listed in Table 3.1, where $A_i$s and $B_i$s are the average number of transitional and transversional substitutions at an $i$th type of site, respectively.

<table>
<thead>
<tr>
<th>parameter</th>
<th>estimator</th>
<th>variance of estimator</th>
</tr>
</thead>
<tbody>
<tr>
<td>$P_S$</td>
<td>$\hat{P}_S = \frac{A_3 + \frac{3}{4} A_2 + \frac{1}{4} A_1}{L_S}$</td>
<td>$\sigma_{SA}^2 = \frac{N_3 P_3 A_3 + \frac{1}{4} N_2 P_2 A_2 + \frac{1}{16} N_1 P_1 A_1}{L_S^2}$</td>
</tr>
<tr>
<td>$Q_S$</td>
<td>$\hat{Q}_S = \frac{B_3}{N_3}$</td>
<td>$\sigma_{SB}^2 = \frac{P_3 B_3}{N_3}$</td>
</tr>
<tr>
<td>$P_A$</td>
<td>$\hat{P}_A = \frac{A_0}{N_0}$</td>
<td>$\sigma_{AA}^2 = \frac{P_0 A_0}{N_0}$</td>
</tr>
<tr>
<td>$Q_A$</td>
<td>$\hat{Q}_A = \frac{B_0 + \frac{3}{4} B_1 + \frac{1}{4} B_2}{L_A}$</td>
<td>$\sigma_{AB}^2 = \frac{N_0 P_0 B_0 + \frac{1}{4} N_1 P_1 B_1 + \frac{1}{16} N_2 P_2 B_2}{L_A^2}$</td>
</tr>
</tbody>
</table>

Table 3.1: Primary Parameters and their Estimators in Two-Parameter Model

At this point it is straightforward to compute the estimators of

$$K_S = -\frac{1}{2} \ln[(1 - 2P_S - Q_S)\sqrt{1 - 2Q_S}]$$

$$K_A = -\frac{1}{2} \ln[(1 - 2P_A - Q_A)\sqrt{1 - 2Q_A}]$$

and their corresponding approximate variances once we realize that $\hat{P}_S$, $\hat{Q}_S$, $\hat{P}_A$, and $\hat{Q}_A$ are the maximum likelihood estimators of $P_S$, $Q_S$, $P_A$, and $Q_A$ respectively:

$$\hat{K}_S = -\frac{1}{2} \ln[(1 - 2\hat{P}_S - \hat{Q}_S)\sqrt{1 - 2\hat{Q}_S}]$$ (3.32)

$$\hat{K}_A = -\frac{1}{2} \ln[(1 - 2\hat{P}_A - \hat{Q}_A)\sqrt{1 - 2\hat{Q}_A}]$$ (3.33)

$$\text{var}(\hat{K}_S) \approx c_3^2 \sigma_{SA}^2 + 2c_3c_4 \text{cov}(\hat{P}_S, \hat{Q}_S) + c_5^2 \sigma_{SB}^2$$ (3.34)

$$\text{var}(\hat{K}_A) \approx c_5^2 \sigma_{AA}^2 + 2c_5c_6 \text{cov}(\hat{P}_A, \hat{Q}_A) + c_5^2 \sigma_{AB}^2$$ (3.35)
where

\[ c_3 = \frac{1}{1 - 2P_S - Q_S} \]
\[ c_4 = \frac{1}{2} \left( \frac{1}{1 - 2P_S - Q_S} + \frac{1}{1 - 2Q_S} \right) \]
\[ c_5 = \frac{1}{1 - 2P_A - Q_A} \]
\[ c_6 = \frac{1}{2} \left( \frac{1}{1 - 2P_A - Q_A} + \frac{1}{1 - 2Q_A} \right) \]
\[ \text{cov}(\hat{P}_S, \hat{Q}_S) = -\frac{1}{L_S} p_3.A p_{3.B} \]
\[ \text{cov}(\hat{P}_A, \hat{Q}_A) = -\frac{1}{L_A} p_{0.A p_{0.B}}. \]

We will now follow the same steps as for the one-parameter model to obtain the variance of the ratio \( \hat{K}_S / \hat{K}_A \). To do so, we also need to find out the covariance between \( \hat{P}_S \) and \( \hat{Q}_A \) first.

\[
\text{cov}(\hat{P}_S, \hat{Q}_A) = \frac{1}{L_S L_A} (\text{cov}(\frac{2}{3} A_2, \frac{1}{3} B_2) + \text{cov}(\frac{1}{3} A_1, \frac{2}{3} B_1))
\]
\[ = -\frac{1}{9 L_S L_A} (N_2 p_{2.A} p_{2.B} + N_1 p_{1.A} p_{1.B}). \]

The final major result of this section is given as follows:

\[
\text{var}(\frac{\hat{K}_S}{\hat{K}_A}) = c_7^2 \sigma_{SA}^2 + c_8^2 \sigma_{SB}^2 + c_9^2 \sigma_{AA}^2 + c_{10}^2 \sigma_{AB}^2 + 2c_7c_8 \text{cov}(\hat{P}_S, \hat{Q}_S) + 2c_7c_{10} \text{cov}(\hat{P}_S, \hat{Q}_A) + 2c_9c_{10} \text{cov}(\hat{P}_A, \hat{Q}_A) \]

(3.36)

where

\[ c_7 = \frac{2}{(1 - 2P_S - Q_S) \ln[(1 - 2P_A - Q_A)\sqrt{1 - 2Q_A}]} \]
\[ c_8 = -\frac{1}{\ln[(1 - 2P_A - Q_A)\sqrt{1 - 2Q_A}]} \left( \frac{1}{1 - 2P_S - Q_S} + \frac{1}{1 - 2Q_S} \right) \]
\[ c_9 = \frac{2 \ln[(1 - 2P_S - Q_S)\sqrt{1 - 2Q}]}{(1 - 2P_A - Q_A)(\ln[(1 - 2P_A - Q_A)\sqrt{1 - 2Q_A}]^2} \]
\[ c_{10} = \frac{\ln[(1 - 2P_S - Q_S)\sqrt{1 - 2Q_S}]}{(\ln[(1 - 2P_A - Q_A)\sqrt{1 - 2Q_A}])^2} \left( \frac{1}{1 - 2P_A - Q_A} + \frac{1}{1 - 2Q_A} \right). \]

To estimate the values of \( K_S, K_A \), the ratio of \( K_S / K_A \), and their corresponding variances, we can simply replace the \( p_{i,j} \)'s and \( q_{i,j} \)'s in the formulas developed in this section by their estimates, namely \( \frac{M_{i,j}}{N_i} \) with the one-parameter model approach, or \( \frac{A_i}{N_i} \) and \( \frac{B_i}{N_i} \) with the two-parameter approach. Here \( i \) is the index of the type of site, \( j \) is the index of synonymous or nonsynonymous mutation, and \( A \) and \( B \) are the indices of transitional or transversional mutation.

### 3.1.3 Modifications

The importance of assessing the variation of a estimator is easily seen. The results developed above are intended to address this matter. For example, variances are needed to construct a confidence interval for the parameter of interest. If we let \( z_{\alpha/2} \) denote the percentage point from the standard normal distribution having a right-tail probability equal to \( \alpha/2 \), a 100(1 - \( \alpha \)) percent confidence interval for an unknown parameter, say \( K_S / K_A \), can be constructed in the standard manner to give \( (\hat{K}_S / \hat{K}_A \pm z_{\alpha/2}\sqrt{\text{var}(\hat{K}_S / \hat{K}_A)}) \). However, we shall be very careful when doing so for two reasons: (1) the parameters \( K_S, K_A \), and \( K_S / K_A \) should all have non-negative values by their biological meaning, while the confidence interval constructed above can not be guaranteed to cover only the positive region; (2) the accuracy of the variance estimates depends heavily on the normality property of the estimators. For large samples, i.e., long DNA sequences, the estimators of \( K_S \) and \( K_A \) will be close to normally distributed. However, the ratio of \( \hat{K}_S / \hat{K}_A \) does not have a symmetric distribution. Thus the convergence of it to the normal distribution is more difficult. Consequently, the inference based on the asymptotic normality has to be cautiously
interpreted, especially in practice, when the typical length is in the range of hundreds of sites.

The standard statistical method to deal with these difficulties in estimating ratios is to consider the log transform instead. The log transform, having an additive rather than a multiplicative structure, converges more rapidly to a normal distribution. Therefore, the confidence interval based on the log transform usually provides a more accurate estimate. By the large-sample normality of $\log(\hat{K}_S/\hat{K}_A)$, an approximate $100(1 - \alpha)$ percent confidence interval for $\log(K_S/K_A)$ can be constructed as follows

$$\log(\hat{K}_S/\hat{K}_A) \pm z_{\alpha/2} \hat{\sigma}(\log(\hat{K}_S/\hat{K}_A)) = \log(\hat{K}_S/\hat{K}_A) \pm z_{\alpha/2} \frac{\hat{\sigma}(\hat{K}_S/\hat{K}_A)}{\hat{K}_S/\hat{K}_A}.$$  

Exponentiating endpoints of this confidence interval gives the suggested confidence interval for $K_S/K_A$

$$\left(\hat{K}_S/\hat{K}_A e^{-z_{\alpha/2} \frac{\hat{\sigma}(\hat{K}_S/\hat{K}_A)}{\hat{K}_S/\hat{K}_A}}, \hat{K}_S/\hat{K}_A e^{z_{\alpha/2} \frac{\hat{\sigma}(\hat{K}_S/\hat{K}_A)}{\hat{K}_S/\hat{K}_A}}\right).$$  

3.2 Simulation Studies

In the previous section, we have proposed a new approach for estimating $\hat{K}_S$, $\hat{K}_A$, $\hat{K}_S/\hat{K}_A$, and their corresponding variances. The estimators were developed by assuming the Bernoulli distribution for seeing a mutation at a particular type of site. One question raised naturally is how accurate the approximations are. Although this question is hard to answer in general, an assessment of how these estimators will perform in the presence of a variety of truths will be a valuable tool. In this section we use a simulation study to help us make these assessments. This simulation study will also compare the results obtained by the proposed method and those given by Li’s [36] method. These two methods differ in two ways: first, in how different types
of sites and mutations occurring at these sites are weighted; second, Li ([36]) makes the multiple-mutation correction for each type of site, while we propose to do that for the entire sequence. The comparisons are made based on the more general two-parameter model.

Six sets of values of $K_S$ and $K_A$ used in the simulation study to generate the nucleotide sequences are summarized in Table 3.2. The data created by the first three sets of parameters will mutate at a lower rate than that by the last three sets. They are the truth we should compare the estimates to. We use these values because actual data suggest that the ratio of $K_S$ and $K_A$ is usually greater than 1, and seldom exceeds 5.

<table>
<thead>
<tr>
<th>parameters</th>
<th>low rate values</th>
<th>high rate values</th>
</tr>
</thead>
<tbody>
<tr>
<td>$K_S$</td>
<td>0.04 0.08 0.2</td>
<td>0.1 0.2 0.5</td>
</tr>
<tr>
<td>$K_A$</td>
<td>0.04 0.04 0.04</td>
<td>0.1 0.1 0.1</td>
</tr>
<tr>
<td>$K_S/K_A$</td>
<td>1 2 5</td>
<td>1 2 5</td>
</tr>
</tbody>
</table>

Table 3.2: True values of the simulation study

Nucleotide sequences of $L=3000$, 900, 480, and 120 sites are used in this study. Nucleotide sequences of 3000 sites are used in order to minimize the effect of stochastic fluctuations due to the small number of sites compared. 900 is thought to be approximately the number of sites in typical genes analyzed in molecular studies. Occasionally, the number of sites analyzed can be much smaller than 900. This is the case especially when we focus on a particular region of genes. Thus, nucleotide sequences of 480 and 120 sites are used in order to examine the effect of the small
number of sites compared. Computer simulations are conducted 1600 times for each combination of $K_S$, $K_A$, and sequence length.

Throughout these simulation studies, "stop" codons will neither be generated nor be mutated to. Only sense codons will be presented. The scheme for creating the data is given by the following. There are in total 61 sense codons. First $L/3$ pseudo-random numbers ranging from 0 to 60 are generated. Then each of these 61 numbers is converted to one of the 61 sense codons to form an ancestral sequence of length $L$. Secondly, the number of mutations, denoted by $n$, is determined. It is drawn from a Poisson distribution with mean $LK_S$, the true number of synonymous substitutions. Then a site is randomly picked from the original sequence. It is afterwards decided which position it should occupy in its codon. For example, if the 4th site of the sequence is chosen, we know it is the first nucleotide of the second codon. Then it will mutate to one of the other three nucleotides with probability $1/3$ if the mutation leads to a synonymous substitution, and with probability $K_A/3K_S$ if the mutation leads to a nonsynonymous substitution. The new sequence is formed by repeating the above procedure $n$ times. Under this mutation scheme, the true synonymous substitution rate and nonsynonymous substitution rates are $K_S$ and $K_A$ respectively. The only parameters being specified in this simulation when generating the nucleotide sequences are $K_S$ and $K_A$. Though the transitional or transversional mutation rates are often confounding with the values of $K_S$ and $K_A$ since transitional mutation tends to lead to synonymous substitutions while transversional mutation tends to lead to nonsynonymous substitutions, there is actually no explicit restriction on the transitional or transversional mutation rates.
By computer simulations, we compare the accuracies of Li's ([36]) method and our proposed method. We examine not only estimators of $K_S$, $K_A$, and the ratio of $K_S/K_A$, denoted by $\hat{K}_S$, $\hat{K}_A$, and $\hat{K}_S/\hat{K}_A$ respectively, but also those of their corresponding variances and confidence intervals when the comparisons are possible. While the performance of $\hat{K}_S$, $\hat{K}_A$, and $\hat{K}_S/\hat{K}_A$ has been the topic of much research, (e.g., Comeron [5], Ina [28]), the estimates of the corresponding standard deviations are also of great importance. This is more difficult to achieve since the true values of these variations are usually unknown. Here, we will treat the sample variances of $\hat{K}_S$, $\hat{K}_A$, and $\hat{K}_S/\hat{K}_A$ as their true variances. This can be argued to be approximately true by the Law of Large Numbers. Then we will compare the estimates of the standard deviations of $\hat{K}_S$, $\hat{K}_A$, and $\hat{K}_S/\hat{K}_A$ (when applicable) to their corresponding "true" values. Since both methods are based on some large sample approximations, the estimators, in general, should provide better estimates as the sequences under study become longer. As pointed out at the beginning of this chapter, all methods that employ one-parameter and two-parameter models can provide appropriate estimations only when the mutation rate between the two sequences is not very high. Otherwise, they will fail to provide the estimates.

Overall, the studies show that the two methods perform similarly, although the new method is typically slightly more accurate if there is any difference. According to the simulation results listed in Tables 3.3, 3.4, and 3.5, $K_S$ is almost always overestimated while $K_A$ is underestimated by both methods. Consequentially, the ratio is always overestimated by both methods. The estimates are usually more accurate when the true parameters \{\$K_S, K_A\} take the bigger sets of values \{0.1, 0.1\}, \{0.2, 0.1\}, or \{0.5, 0.1\}, especially for the shorter sequences of $L = 120$ or 480. The
<table>
<thead>
<tr>
<th></th>
<th>Expectation</th>
<th>Li</th>
<th>New</th>
<th></th>
<th>Expectation</th>
<th>Li</th>
<th>New</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>L=120</td>
<td></td>
<td></td>
<td></td>
<td>L=120</td>
<td></td>
<td></td>
</tr>
<tr>
<td>n</td>
<td>1600</td>
<td>1549</td>
<td>1550</td>
<td>1600</td>
<td>1548</td>
<td>1581</td>
<td></td>
</tr>
<tr>
<td>( \hat{K}_S )</td>
<td>0.04</td>
<td>0.061</td>
<td>0.049</td>
<td>0.1</td>
<td>0.103</td>
<td>0.100</td>
<td></td>
</tr>
<tr>
<td>SD of ( \hat{K}_S )</td>
<td>0.041</td>
<td>0.051</td>
<td>0.051</td>
<td>0.082</td>
<td>0.088</td>
<td>0.087</td>
<td></td>
</tr>
<tr>
<td>( \hat{K}_A )</td>
<td>0.04</td>
<td>0.036</td>
<td>0.036</td>
<td>0.1</td>
<td>0.087</td>
<td>0.088</td>
<td></td>
</tr>
<tr>
<td>SD of ( \hat{K}_A )</td>
<td>0.010</td>
<td>0.020</td>
<td>0.020</td>
<td>0.026</td>
<td>0.042</td>
<td>0.043</td>
<td></td>
</tr>
<tr>
<td>( \hat{K}_S/\hat{K}_A )</td>
<td>1</td>
<td>1.640</td>
<td>1.564</td>
<td>1</td>
<td>1.245</td>
<td>1.222</td>
<td></td>
</tr>
<tr>
<td>SD of ( \hat{K}_S/\hat{K}_A )</td>
<td>1.768</td>
<td>NA</td>
<td>1.390</td>
<td>0.861</td>
<td>NA</td>
<td>0.805</td>
<td></td>
</tr>
<tr>
<td></td>
<td>L=480</td>
<td></td>
<td></td>
<td></td>
<td>L=480</td>
<td></td>
<td></td>
</tr>
<tr>
<td>n</td>
<td>1600</td>
<td>1585</td>
<td>1585</td>
<td>1600</td>
<td>1599</td>
<td>1599</td>
<td></td>
</tr>
<tr>
<td>( \hat{K}_S )</td>
<td>0.04</td>
<td>0.045</td>
<td>0.045</td>
<td>0.1</td>
<td>0.102</td>
<td>0.099</td>
<td></td>
</tr>
<tr>
<td>SD of ( \hat{K}_S )</td>
<td>0.023</td>
<td>0.021</td>
<td>0.021</td>
<td>0.037</td>
<td>0.034</td>
<td>0.034</td>
<td></td>
</tr>
<tr>
<td>( \hat{K}_A )</td>
<td>0.04</td>
<td>0.041</td>
<td>0.041</td>
<td>0.1</td>
<td>0.091</td>
<td>0.092</td>
<td></td>
</tr>
<tr>
<td>SD of ( \hat{K}_A )</td>
<td>0.014</td>
<td>0.011</td>
<td>0.011</td>
<td>0.020</td>
<td>0.017</td>
<td>0.017</td>
<td></td>
</tr>
<tr>
<td>( \hat{K}_S/\hat{K}_A )</td>
<td>1</td>
<td>1.231</td>
<td>1.217</td>
<td>1</td>
<td>1.164</td>
<td>1.111</td>
<td></td>
</tr>
<tr>
<td>SD of ( \hat{K}_S/\hat{K}_A )</td>
<td>0.857</td>
<td>NA</td>
<td>0.795</td>
<td>0.436</td>
<td>NA</td>
<td>0.441</td>
<td></td>
</tr>
<tr>
<td></td>
<td>L=900</td>
<td></td>
<td></td>
<td></td>
<td>L=900</td>
<td></td>
<td></td>
</tr>
<tr>
<td>n</td>
<td>1600</td>
<td>1600</td>
<td>1600</td>
<td>1600</td>
<td>1600</td>
<td>1600</td>
<td></td>
</tr>
<tr>
<td>( \hat{K}_S )</td>
<td>0.04</td>
<td>0.041</td>
<td>0.041</td>
<td>0.1</td>
<td>0.104</td>
<td>0.101</td>
<td></td>
</tr>
<tr>
<td>SD of ( \hat{K}_S )</td>
<td>0.013</td>
<td>0.014</td>
<td>0.014</td>
<td>0.024</td>
<td>0.025</td>
<td>0.025</td>
<td></td>
</tr>
<tr>
<td>( \hat{K}_A )</td>
<td>0.04</td>
<td>0.037</td>
<td>0.037</td>
<td>0.1</td>
<td>0.094</td>
<td>0.096</td>
<td></td>
</tr>
<tr>
<td>SD of ( \hat{K}_A )</td>
<td>0.006</td>
<td>0.007</td>
<td>0.007</td>
<td>0.009</td>
<td>0.013</td>
<td>0.013</td>
<td></td>
</tr>
<tr>
<td>( \hat{K}_S/\hat{K}_A )</td>
<td>1</td>
<td>1.173</td>
<td>1.127</td>
<td>1</td>
<td>1.130</td>
<td>1.079</td>
<td></td>
</tr>
<tr>
<td>SD of ( \hat{K}_S/\hat{K}_A )</td>
<td>0.496</td>
<td>NA</td>
<td>0.503</td>
<td>0.306</td>
<td>NA</td>
<td>0.303</td>
<td></td>
</tr>
<tr>
<td></td>
<td>L=3000</td>
<td></td>
<td></td>
<td></td>
<td>L=3000</td>
<td></td>
<td></td>
</tr>
<tr>
<td>n</td>
<td>1600</td>
<td>1600</td>
<td>1600</td>
<td>1600</td>
<td>1600</td>
<td>1600</td>
<td></td>
</tr>
<tr>
<td>( \hat{K}_S )</td>
<td>0.04</td>
<td>0.041</td>
<td>0.040</td>
<td>0.1</td>
<td>0.100</td>
<td>0.100</td>
<td></td>
</tr>
<tr>
<td>SD of ( \hat{K}_S )</td>
<td>0.010</td>
<td>0.008</td>
<td>0.008</td>
<td>0.012</td>
<td>0.013</td>
<td>0.013</td>
<td></td>
</tr>
<tr>
<td>( \hat{K}_A )</td>
<td>0.04</td>
<td>0.038</td>
<td>0.038</td>
<td>0.1</td>
<td>0.094</td>
<td>0.096</td>
<td></td>
</tr>
<tr>
<td>SD of ( \hat{K}_A )</td>
<td>0.008</td>
<td>0.007</td>
<td>0.008</td>
<td>0.006</td>
<td>0.007</td>
<td>0.007</td>
<td></td>
</tr>
<tr>
<td>( \hat{K}_S/\hat{K}_A )</td>
<td>1</td>
<td>1.108</td>
<td>1.080</td>
<td>1</td>
<td>1.109</td>
<td>1.078</td>
<td></td>
</tr>
<tr>
<td>SD of ( \hat{K}_S/\hat{K}_A )</td>
<td>0.255</td>
<td>NA</td>
<td>0.257</td>
<td>0.164</td>
<td>NA</td>
<td>0.166</td>
<td></td>
</tr>
</tbody>
</table>

Table 3.3: Mean and standard deviations of \( \hat{K}_S \), \( \hat{K}_A \), and \( \hat{K}_S/\hat{K}_A \) obtained when the ratio is 1 by excluding inapplicable cases. \( n=\) number of applicable cases.
<table>
<thead>
<tr>
<th></th>
<th>Expectation</th>
<th>Li</th>
<th>New</th>
<th>Expectation</th>
<th>Li</th>
<th>New</th>
</tr>
</thead>
<tbody>
<tr>
<td>$L=120$</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>$n$</td>
<td>1600</td>
<td>1538</td>
<td>1554</td>
<td>1600</td>
<td>1572</td>
<td>1575</td>
</tr>
<tr>
<td>$\hat{K}_S$</td>
<td>0.08</td>
<td>0.093</td>
<td>0.091</td>
<td>0.2</td>
<td>0.207</td>
<td>0.202</td>
</tr>
<tr>
<td>SD of $\hat{K}_S$</td>
<td>0.072</td>
<td>0.075</td>
<td>0.074</td>
<td>0.131</td>
<td>0.113</td>
<td>0.109</td>
</tr>
<tr>
<td>$\hat{K}_A$</td>
<td>0.04</td>
<td>0.034</td>
<td>0.035</td>
<td>0.1</td>
<td>0.084</td>
<td>0.086</td>
</tr>
<tr>
<td>SD of $\hat{K}_A$</td>
<td>0.026</td>
<td>0.026</td>
<td>0.026</td>
<td>0.041</td>
<td>0.032</td>
<td>0.031</td>
</tr>
<tr>
<td>$\hat{K}_S/\hat{K}_A$</td>
<td>2</td>
<td>3.014</td>
<td>2.906</td>
<td>2</td>
<td>2.954</td>
<td>2.725</td>
</tr>
<tr>
<td>SD of $\hat{K}_S/\hat{K}_A$</td>
<td>2.799</td>
<td>NA</td>
<td>2.416</td>
<td>2.563</td>
<td>NA</td>
<td>2.066</td>
</tr>
<tr>
<td>$L=480$</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>$n$</td>
<td>1600</td>
<td>1582</td>
<td>1591</td>
<td>1600</td>
<td>1598</td>
<td>1600</td>
</tr>
<tr>
<td>$\hat{K}_S$</td>
<td>0.08</td>
<td>0.085</td>
<td>0.083</td>
<td>0.2</td>
<td>0.204</td>
<td>0.201</td>
</tr>
<tr>
<td>SD of $\hat{K}_S$</td>
<td>0.061</td>
<td>0.063</td>
<td>0.062</td>
<td>0.049</td>
<td>0.047</td>
<td>0.047</td>
</tr>
<tr>
<td>$\hat{K}_A$</td>
<td>0.04</td>
<td>0.036</td>
<td>0.038</td>
<td>0.1</td>
<td>0.091</td>
<td>0.093</td>
</tr>
<tr>
<td>SD of $\hat{K}_A$</td>
<td>0.018</td>
<td>0.023</td>
<td>0.022</td>
<td>0.017</td>
<td>0.016</td>
<td>0.016</td>
</tr>
<tr>
<td>$\hat{K}_S/\hat{K}_A$</td>
<td>2</td>
<td>2.395</td>
<td>2.292</td>
<td>2</td>
<td>2.208</td>
<td>2.195</td>
</tr>
<tr>
<td>SD of $\hat{K}_S/\hat{K}_A$</td>
<td>1.270</td>
<td>NA</td>
<td>1.327</td>
<td>0.717</td>
<td>NA</td>
<td>0.714</td>
</tr>
<tr>
<td>$L=900$</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>$n$</td>
<td>1600</td>
<td>1600</td>
<td>1600</td>
<td>1600</td>
<td>1600</td>
<td>1600</td>
</tr>
<tr>
<td>$\hat{K}_S$</td>
<td>0.08</td>
<td>0.084</td>
<td>0.082</td>
<td>0.2</td>
<td>0.204</td>
<td>0.203</td>
</tr>
<tr>
<td>SD of $\hat{K}_S$</td>
<td>0.019</td>
<td>0.020</td>
<td>0.020</td>
<td>0.035</td>
<td>0.036</td>
<td>0.037</td>
</tr>
<tr>
<td>$\hat{K}_A$</td>
<td>0.04</td>
<td>0.038</td>
<td>0.039</td>
<td>0.1</td>
<td>0.094</td>
<td>0.094</td>
</tr>
<tr>
<td>SD of $\hat{K}_A$</td>
<td>0.006</td>
<td>0.007</td>
<td>0.007</td>
<td>0.011</td>
<td>0.012</td>
<td>0.012</td>
</tr>
<tr>
<td>$\hat{K}_S/\hat{K}_A$</td>
<td>2</td>
<td>2.177</td>
<td>2.138</td>
<td>2</td>
<td>2.187</td>
<td>2.094</td>
</tr>
<tr>
<td>SD of $\hat{K}_S/\hat{K}_A$</td>
<td>0.826</td>
<td>NA</td>
<td>0.807</td>
<td>0.509</td>
<td>NA</td>
<td>0.507</td>
</tr>
<tr>
<td>$L=3000$</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>$n$</td>
<td>1600</td>
<td>1600</td>
<td>1600</td>
<td>1600</td>
<td>1600</td>
<td>1600</td>
</tr>
<tr>
<td>$\hat{K}_S$</td>
<td>0.08</td>
<td>0.083</td>
<td>0.082</td>
<td>0.2</td>
<td>0.201</td>
<td>0.200</td>
</tr>
<tr>
<td>SD of $\hat{K}_S$</td>
<td>0.012</td>
<td>0.012</td>
<td>0.012</td>
<td>0.018</td>
<td>0.019</td>
<td>0.019</td>
</tr>
<tr>
<td>$\hat{K}_A$</td>
<td>0.04</td>
<td>0.038</td>
<td>0.039</td>
<td>0.1</td>
<td>0.096</td>
<td>0.097</td>
</tr>
<tr>
<td>SD of $\hat{K}_A$</td>
<td>0.004</td>
<td>0.004</td>
<td>0.004</td>
<td>0.006</td>
<td>0.007</td>
<td>0.007</td>
</tr>
<tr>
<td>$\hat{K}_S/\hat{K}_A$</td>
<td>2</td>
<td>2.196</td>
<td>2.147</td>
<td>2</td>
<td>2.149</td>
<td>2.082</td>
</tr>
<tr>
<td>SD of $\hat{K}_S/\hat{K}_A$</td>
<td>0.403</td>
<td>NA</td>
<td>0.399</td>
<td>0.257</td>
<td>NA</td>
<td>0.253</td>
</tr>
</tbody>
</table>

Table 3.4: Mean and standard deviations of $\hat{K}_S$, $\hat{K}_A$, and $\hat{K}_S/\hat{K}_A$ obtained when the ratio is 2 by excluding inapplicable cases. $n=$number of applicable cases.
<table>
<thead>
<tr>
<th></th>
<th>Expectation</th>
<th>Li</th>
<th>New</th>
<th>Expectation</th>
<th>Li</th>
<th>New</th>
</tr>
</thead>
<tbody>
<tr>
<td>$L=120$</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>$n$</td>
<td>1600</td>
<td>1548</td>
<td>1567</td>
<td>1600</td>
<td>1504</td>
<td>1594</td>
</tr>
<tr>
<td>$\hat{K}_S$</td>
<td>0.2</td>
<td>0.218</td>
<td>0.207</td>
<td>0.5</td>
<td>0.575</td>
<td>0.544</td>
</tr>
<tr>
<td>SD of $\hat{K}_S$</td>
<td>0.118</td>
<td>0.117</td>
<td>0.113</td>
<td>0.208</td>
<td>0.231</td>
<td>0.215</td>
</tr>
<tr>
<td>$\hat{K}_A$</td>
<td>0.04</td>
<td>0.034</td>
<td>0.035</td>
<td>0.1</td>
<td>0.098</td>
<td>0.102</td>
</tr>
<tr>
<td>SD of $\hat{K}_A$</td>
<td>0.022</td>
<td>0.022</td>
<td>0.022</td>
<td>0.035</td>
<td>0.036</td>
<td>0.035</td>
</tr>
<tr>
<td>$\hat{K}_S/\hat{K}_A$</td>
<td>5</td>
<td>6.329</td>
<td>6.121</td>
<td>5</td>
<td>6.131</td>
<td>5.732</td>
</tr>
<tr>
<td>SD of $\hat{K}_S/\hat{K}_A$</td>
<td>6.141</td>
<td>NA</td>
<td>6.127</td>
<td>4.238</td>
<td>NA</td>
<td>3.972</td>
</tr>
<tr>
<td>$L=480$</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>$n$</td>
<td>1600</td>
<td>1600</td>
<td>1600</td>
<td>1600</td>
<td>1593</td>
<td>1600</td>
</tr>
<tr>
<td>$\hat{K}_S$</td>
<td>0.2</td>
<td>0.211</td>
<td>0.204</td>
<td>0.5</td>
<td>0.507</td>
<td>0.503</td>
</tr>
<tr>
<td>SD of $\hat{K}_S$</td>
<td>0.044</td>
<td>0.046</td>
<td>0.046</td>
<td>0.092</td>
<td>0.095</td>
<td>0.095</td>
</tr>
<tr>
<td>$\hat{K}_A$</td>
<td>0.04</td>
<td>0.036</td>
<td>0.037</td>
<td>0.1</td>
<td>0.096</td>
<td>0.097</td>
</tr>
<tr>
<td>SD of $\hat{K}_A$</td>
<td>0.010</td>
<td>0.010</td>
<td>0.010</td>
<td>0.017</td>
<td>0.017</td>
<td>0.017</td>
</tr>
<tr>
<td>$\hat{K}_S/\hat{K}_A$</td>
<td>5</td>
<td>5.642</td>
<td>5.311</td>
<td>5</td>
<td>5.428</td>
<td>5.276</td>
</tr>
<tr>
<td>SD of $\hat{K}_S/\hat{K}_A$</td>
<td>2.342</td>
<td>NA</td>
<td>2.101</td>
<td>1.436</td>
<td>NA</td>
<td>1.412</td>
</tr>
<tr>
<td>$L=900$</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>$n$</td>
<td>1600</td>
<td>1600</td>
<td>1600</td>
<td>1600</td>
<td>1600</td>
<td>1600</td>
</tr>
<tr>
<td>$\hat{K}_S$</td>
<td>0.2</td>
<td>0.201</td>
<td>0.201</td>
<td>0.5</td>
<td>0.506</td>
<td>0.501</td>
</tr>
<tr>
<td>SD of $\hat{K}_S$</td>
<td>0.034</td>
<td>0.035</td>
<td>0.035</td>
<td>0.066</td>
<td>0.068</td>
<td>0.069</td>
</tr>
<tr>
<td>$\hat{K}_A$</td>
<td>0.04</td>
<td>0.037</td>
<td>0.038</td>
<td>0.1</td>
<td>0.094</td>
<td>0.095</td>
</tr>
<tr>
<td>SD of $\hat{K}_A$</td>
<td>0.007</td>
<td>0.008</td>
<td>0.008</td>
<td>0.012</td>
<td>0.012</td>
<td>0.012</td>
</tr>
<tr>
<td>$\hat{K}_S/\hat{K}_A$</td>
<td>5</td>
<td>5.369</td>
<td>5.171</td>
<td>5</td>
<td>5.267</td>
<td>5.105</td>
</tr>
<tr>
<td>SD of $\hat{K}_S/\hat{K}_A$</td>
<td>1.490</td>
<td>NA</td>
<td>1.461</td>
<td>1.052</td>
<td>NA</td>
<td>1.004</td>
</tr>
<tr>
<td>$L=3000$</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>$n$</td>
<td>1600</td>
<td>1600</td>
<td>1600</td>
<td>1600</td>
<td>1600</td>
<td>1600</td>
</tr>
<tr>
<td>$\hat{K}_S$</td>
<td>0.2</td>
<td>0.205</td>
<td>0.201</td>
<td>0.5</td>
<td>0.508</td>
<td>0.501</td>
</tr>
<tr>
<td>SD of $\hat{K}_S$</td>
<td>0.019</td>
<td>0.020</td>
<td>0.020</td>
<td>0.039</td>
<td>0.037</td>
<td>0.038</td>
</tr>
<tr>
<td>$\hat{K}_A$</td>
<td>0.04</td>
<td>0.038</td>
<td>0.039</td>
<td>0.1</td>
<td>0.096</td>
<td>0.097</td>
</tr>
<tr>
<td>SD of $\hat{K}_A$</td>
<td>0.004</td>
<td>0.004</td>
<td>0.004</td>
<td>0.007</td>
<td>0.007</td>
<td>0.007</td>
</tr>
<tr>
<td>$\hat{K}_S/\hat{K}_A$</td>
<td>5</td>
<td>5.179</td>
<td>5.002</td>
<td>5</td>
<td>5.106</td>
<td>5.008</td>
</tr>
<tr>
<td>SD of $\hat{K}_S/\hat{K}_A$</td>
<td>0.749</td>
<td>NA</td>
<td>0.759</td>
<td>0.528</td>
<td>NA</td>
<td>0.530</td>
</tr>
</tbody>
</table>

Table 3.5: Mean and standard deviations of $\hat{K}_S$, $\hat{K}_A$, and $\hat{K}_S/\hat{K}_A$ obtained when the ratio is 5 by excluding inapplicable cases. $n=$number of applicable cases
sequence length has significant influence on the estimates. As expected, all estimators perform better when the sequences become longer. One thing we notice from the study is that sometimes when the average simulated values of $K_S$ and $K_A$ are very close to their true values, the average of $K_S/K_A$ surprisingly still overestimates the ratio of $K_S$ to $K_A$ by a substantial margin. The reason for causing this confusion can be seen from the following example:

Suppose after repeating the simulation three times, $K_S$ and $K_A$ are estimated to be \{0.3, 0.2, 0.2\} and \{0.03, 0.04, 0.05\}, respectively. Then the average of $K_S$ is about 0.23, and the average of $K_A$ is 0.04. Both fall inside the ranges of the observed values of $K_S$ and $K_A$ in Table 3.5 when $K_S = 0.2$ and $K_A = 0.04$. However, the average ratio varies from $(0.3/0.05 + 0.2/0.04 + 0.2/0.03)/3 \approx 5.9$ to $(0.3/0.03 + 0.2/0.04 + 0.2/0.05)/3 \approx 6.3$, which is always much higher than the expected value of 5. In other words, even though the sample means of $K_S$ and $K_A$ are close to the expected values of $K_S$ and $K_A$ respectively, the sample mean of $K_S/K_A$ still might overestimate the expected value of $K_S/K_A$ because of the ratio structure.

Another reason for the seemingly biased estimation is due to the long right tail shaped distribution of $K_S/K_A$. The sample median therefore may be better representative of a typical value than the sample mean that is reported. For example, the sample medians for the case of both $K_S$ and $K_A$ equal to 0.04 based on the new method are 1.317, 1.112, 1.078, and 1.063 for the four sample sizes employed in the simulation study. The bias in the median is about half of the bias in the simulated sample means listed in Table 3.3, especially for the smaller sample size.

Overall, the simulations also show that both methods provide good estimates for the variances of $K_S$ and $K_A$. The estimates for the standard deviations of $K_S/K_A$
by our method are also reasonably close to the true values, while Li's ([36]) method does not provide an estimate of the standard deviation of $\hat{K}_S/\hat{K}_A$. Both the true and estimated standard deviations of $\hat{K}_S$, $\hat{K}_A$, and $\hat{K}_S/\hat{K}_A$ are decreasing as the length of sequence increases as expected. Interestingly, the standard deviations tend to be larger for the data set generated with smaller values of $K_S$ and $K_A$. Combining this with the previous finding that $\hat{K}_S$, $\hat{K}_A$, and $\hat{K}_S/\hat{K}_A$ all provide better estimates for larger values of $K_S$ and $K_A$, the accuracies of the confidence intervals are understandingably robust with respect to the variation of $K_S$ and $K_A$. This is illustrated by Tables 3.6, 3.7, and 3.8.

The entries of these three tables are the percentages of 95% confidence intervals that actually cover the true values. Ideally, 95% of the confidence intervals should cover the true values. Hence, the ones with coverage percentages closer to 95% are considered to be more accurate. In general, the suggested intervals based on the log transform are better than the original ones for the parameter of $K_S/K_A$ when the sequences under study are not bigger than 480. Here "better" is illustrated by the fact that the percentages of intervals containing the true values are closer to the expected 95%. However, the confidence intervals constructed by both methods provide similar estimations when the sequence length exceeds 480. This may be explained by the fact that the construction of the confidence intervals by both methods heavily depends on the normal structure of the estimators. The estimator of $K_S/K_A$ itself usually converges to the normal distribution at a much slower rate than that of the log transformed estimator. When the sample size is small, the confidence intervals constructed by log transformation provide a significantly better estimation than the original ones. But when the sequences are sufficiently long, the distribution of both
<table>
<thead>
<tr>
<th>Parameter Value</th>
<th>Li</th>
<th>New</th>
<th>Parameter Value</th>
<th>Li</th>
<th>New</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>$K_S=0.04$</td>
<td>94.2</td>
<td>96.7</td>
<td>$K_S=0.1$</td>
<td>96.3</td>
<td>96.1</td>
</tr>
<tr>
<td>$K_A=0.04$</td>
<td>97.0</td>
<td>96.4</td>
<td>$K_A=0.1$</td>
<td>99.7</td>
<td>99.6</td>
</tr>
<tr>
<td>$K_S/K_A=1$</td>
<td>NA</td>
<td>87.6</td>
<td>$K_S/K_A=1$</td>
<td>NA</td>
<td>90.0</td>
</tr>
<tr>
<td>$\log(K_S/K_A) = 0$</td>
<td>NA</td>
<td>96.3</td>
<td>$\log(K_S/K_A) = 0$</td>
<td>NA</td>
<td>93.6</td>
</tr>
<tr>
<td>$L=480$</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>$K_S=0.04$</td>
<td>90.2</td>
<td>90.3</td>
<td>$K_S=0.1$</td>
<td>92.4</td>
<td>93.1</td>
</tr>
<tr>
<td>$K_A=0.04$</td>
<td>93.5</td>
<td>94.4</td>
<td>$K_A=0.1$</td>
<td>92.9</td>
<td>93.3</td>
</tr>
<tr>
<td>$K_S/K_A=1$</td>
<td>NA</td>
<td>91.8</td>
<td>$K_S/K_A=1$</td>
<td>NA</td>
<td>94.4</td>
</tr>
<tr>
<td>$\log(K_S/K_A) = 0$</td>
<td>NA</td>
<td>94.6</td>
<td>$\log(K_S/K_A) = 0$</td>
<td>NA</td>
<td>95.8</td>
</tr>
<tr>
<td>$L=900$</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>$K_S=0.04$</td>
<td>92.6</td>
<td>93.8</td>
<td>$K_S=0.1$</td>
<td>95.9</td>
<td>95.4</td>
</tr>
<tr>
<td>$K_A=0.04$</td>
<td>90.4</td>
<td>91.5</td>
<td>$K_A=0.1$</td>
<td>95.8</td>
<td>96.6</td>
</tr>
<tr>
<td>$K_S/K_A=1$</td>
<td>NA</td>
<td>93.8</td>
<td>$K_S/K_A=1$</td>
<td>NA</td>
<td>95.6</td>
</tr>
<tr>
<td>$\log(K_S/K_A) = 0$</td>
<td>NA</td>
<td>96.0</td>
<td>$\log(K_S/K_A) = 0$</td>
<td>NA</td>
<td>94.5</td>
</tr>
<tr>
<td>$L=3000$</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>$K_S=0.04$</td>
<td>93.6</td>
<td>94.2</td>
<td>$K_S=0.1$</td>
<td>95.7</td>
<td>95.2</td>
</tr>
<tr>
<td>$K_A=0.04$</td>
<td>93.3</td>
<td>93.5</td>
<td>$K_A=0.1$</td>
<td>92.2</td>
<td>93.4</td>
</tr>
<tr>
<td>$K_S/K_A=1$</td>
<td>NA</td>
<td>96.3</td>
<td>$K_S/K_A=1$</td>
<td>NA</td>
<td>94.8</td>
</tr>
<tr>
<td>$\log(K_S/K_A) = 0$</td>
<td>NA</td>
<td>94.2</td>
<td>$\log(K_S/K_A) = 0$</td>
<td>NA</td>
<td>93.1</td>
</tr>
</tbody>
</table>

Table 3.6: Percentage of confidence intervals containing true values when ratio is 1. The confidence intervals for the log($K_S/K_A$) value are the same as the suggested intervals using log transform for the original parameter $K_S/K_A$.

estimators are very close to the normal distribution, and the confidence intervals constructed by both approaches perform similarly.

Overall, the study shows that both methods tend to overestimate $K_S$ and underestimate $K_A$. This result is similar to the ones observed by Ina ([28]) and many others. One interesting finding is that when the sequence mutates at a higher rate, i.e., the parameters of $\{K_S, K_A\}$ are further away from the extreme value 0, the estimators generally provide more accurate and reliable estimates. The simulation
Table 3.7: Percentages of confidence intervals containing true values when ratio is 2. The confidence intervals for the log($K_S/K_A$) value are the same as the suggested intervals using log transform for the original parameter $K_S/K_A$.

Also shows that $K_S/K_A$ always overestimates $K_S/K_A$ because of the overestimation of $K_S$, underestimation of $K_A$, and its ratio structure. However, the confidence intervals constructed are shown to provide reasonably good estimates of $K_S$, $K_A$, and $K_S/K_A$ even though the point estimators of these parameters appear to be biased.
Table 3.8: Percentages of confidence intervals containing true values when ratio is 5. The confidence intervals for the log($K_S/K_A$) value are the same as the suggested intervals using log transform for the original parameter $K_S/K_A$.

<table>
<thead>
<tr>
<th>Parameter Value</th>
<th>Li New</th>
<th>Parameter Value</th>
<th>Li New</th>
</tr>
</thead>
<tbody>
<tr>
<td>$K_S=0.2$</td>
<td>92.4</td>
<td>92.1</td>
<td>$K_S=0.5$</td>
</tr>
<tr>
<td>$K_A=0.04$</td>
<td>91.0</td>
<td>91.2</td>
<td>$K_A=0.1$</td>
</tr>
<tr>
<td>$K_S/K_A=5$</td>
<td>NA</td>
<td>86.7</td>
<td>$K_S/K_A=5$</td>
</tr>
<tr>
<td>log($K_S/K_A$) = log 5</td>
<td>NA</td>
<td>95.6</td>
<td>log($K_S/K_A$) = log 5</td>
</tr>
<tr>
<td>$L=480$</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>$K_S=0.2$</td>
<td>91.6</td>
<td>90.5</td>
<td>$K_S=0.5$</td>
</tr>
<tr>
<td>$K_A=0.04$</td>
<td>90.6</td>
<td>92.1</td>
<td>$K_A=0.1$</td>
</tr>
<tr>
<td>$K_S/K_A=5$</td>
<td>NA</td>
<td>91.6</td>
<td>$K_S/K_A=5$</td>
</tr>
<tr>
<td>log($K_S/K_A$) = log 5</td>
<td>NA</td>
<td>95.3</td>
<td>log($K_S/K_A$) = log 5</td>
</tr>
<tr>
<td>$L=900$</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>$K_S=0.2$</td>
<td>91.6</td>
<td>90.4</td>
<td>$K_S=0.5$</td>
</tr>
<tr>
<td>$K_A=0.04$</td>
<td>92.1</td>
<td>92.9</td>
<td>$K_A=0.1$</td>
</tr>
<tr>
<td>$K_S/K_A=5$</td>
<td>NA</td>
<td>94.1</td>
<td>$K_S/K_A=5$</td>
</tr>
<tr>
<td>log($K_S/K_A$) = log 5</td>
<td>NA</td>
<td>94.9</td>
<td>log($K_S/K_A$) = log 5</td>
</tr>
<tr>
<td>$L=3000$</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>$K_S=0.2$</td>
<td>94.0</td>
<td>92.3</td>
<td>$K_S=0.5$</td>
</tr>
<tr>
<td>$K_A=0.04$</td>
<td>92.6</td>
<td>93.9</td>
<td>$K_A=0.1$</td>
</tr>
<tr>
<td>$K_S/K_A=5$</td>
<td>NA</td>
<td>94.9</td>
<td>$K_S/K_A=5$</td>
</tr>
<tr>
<td>log($K_S/K_A$) = log 5</td>
<td>NA</td>
<td>95.4</td>
<td>log($K_S/K_A$) = log 5</td>
</tr>
</tbody>
</table>

3.3 Applications

In this section, we will employ the method proposed in the previous section to analyze two sets of nucleotide data.

We first use our method to compare the coding and noncoding regions of the rp49 gene of the closely related *D.pseudoobscura* and *D.subobscura*. Comparison of the rp49 noncoding regions has allowed the study of how molecular evolution has operated in these noncoding regions. Moreover, estimates of the rate of nucleotide substitutions in
the rp49 coding regions in the comparison of these species have allowed confirmation of low rates of silent and nonsilent divergence for the rp49 region and a rough estimate of the time of divergence between *D.pseudoobscura* and *D.subobscura* (Segarra and Aguade [53]).

This example acts as a further test to determine the usefulness and efficacy of our method. The purpose of analyzing this data set is first to compare the estimations of *K*<sub>S</sub> and *K*<sub>A</sub> provided by our method to those by other methods. Secondly, it is thought that, for all functional proteins of the coding region, *K*<sub>S</sub> is higher than *K*<sub>A</sub>, only in the case of pseudogenes of the noncoding region does *K*<sub>S</sub> approach *K*<sub>A</sub>. In this section, we compare both the coding and noncoding regions from the same pair of species, whose *K*<sub>S</sub>/*K*<sub>A</sub> ratios can be considered as known because of the biological functionality. Therefore we will see if the estimates of these ratios are able to confirm this hypothesis. The result is listed in Table 3.9.

The estimates of *K*<sub>S</sub> and *K*<sub>A</sub> for the coding region are lower and higher, respectively, than those obtained by using Li et al.'s ([39]) method as expected since their method tends to overestimate *K*<sub>S</sub> and underestimate *K*<sub>A</sub> as pointed out by Li([36]). The results are close to those obtained by Comeron ([5]). To further explore the proposed method, three different approaches are employed for both the coding and noncoding regions. First, the estimation for both regions is computed for the two original aligned sequences. Then the first nucleotide of the coding and noncoding regions from both sequences are removed and estimation is again computed based on the newly formed sequences. Last, an additional nucleotide is removed from both regions and the same process is repeated. Each of these deletions will cause a shift in the sequential groupings of the nucleotides for forming the codons. If such a shift
occurs in the coding region, \( K_S \) and \( K_A \) will change dramatically because of the biological functional constraint. But if a shift occurs in the noncoding region, they should remain approximately the same.

Table 3.9 indeed coincides with this theoretical conclusion. According to the proposed method, \( \hat{K}_S, \hat{K}_A \) and \( \hat{K}_S/\hat{K}_A \) are substantially different for the coding region and are reasonably close for the noncoding region across the three schemes of estimation. It is also important to look at the estimates of the \( K_S/K_A \) ratio. As stated previously, in the coding region, the synonymous substitution rate should be higher than the nonsynonymous substitution rate, i.e., \( K_S/K_A > 1 \), while for the noncoding region, this ratio should be close to 1. This is exactly what Table 3.9 shows. Importantly, according to the confidence intervals listed in Table 3.9, the \( \hat{K}_S/\hat{K}_A \) ratio is

<table>
<thead>
<tr>
<th>region</th>
<th>( \hat{K}_S )</th>
<th>( \hat{K}_A )</th>
<th>( \hat{K}_S/\hat{K}_A )</th>
<th>SD(( \hat{K}_S/\hat{K}_A ))</th>
<th>95% C.I.</th>
</tr>
</thead>
<tbody>
<tr>
<td>coding region</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>(402 sites)</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Li et al.</td>
<td>0.1769</td>
<td>0.0032</td>
<td>55.281</td>
<td>NA</td>
<td></td>
</tr>
<tr>
<td>Comeron proposed</td>
<td>0.1198</td>
<td>0.0038</td>
<td>31.526</td>
<td>NA</td>
<td></td>
</tr>
<tr>
<td>original</td>
<td>0.1223</td>
<td>0.0038</td>
<td>31.800</td>
<td>33.373</td>
<td>(4.07, 248.73)</td>
</tr>
<tr>
<td>frame shift by 1 site</td>
<td>0.0203</td>
<td>0.0499</td>
<td>0.4067</td>
<td>0.314</td>
<td>(0.09, 1.84)</td>
</tr>
<tr>
<td>frame shift by 2 sites</td>
<td>0.0086</td>
<td>0.0513</td>
<td>0.1673</td>
<td>0.115</td>
<td>(0.04, 0.65)</td>
</tr>
<tr>
<td>noncoding region</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>(996 sites)</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>original</td>
<td>0.1775</td>
<td>0.1607</td>
<td>1.105</td>
<td>0.271</td>
<td>(0.68, 1.79)</td>
</tr>
<tr>
<td>frame shift by 1 site</td>
<td>0.1563</td>
<td>0.1559</td>
<td>1.003</td>
<td>0.251</td>
<td>(0.61, 1.64)</td>
</tr>
<tr>
<td>frame shift by 2 sites</td>
<td>0.2056</td>
<td>0.1646</td>
<td>1.249</td>
<td>0.312</td>
<td>(0.77, 2.04)</td>
</tr>
</tbody>
</table>

Table 3.9: Comparisons of \( D. pseudoobscura \) and \( D. subobscura \). The suggested confidence intervals for the \( K_S/K_A \) ratio are based on the log transform.
significantly higher than 1 for the coding region and close to 1 for all the noncoding regions as expected. This example clearly indicates that our method is not only very comparable to other existing methods for estimating $K_S$, $K_A$, and $K_S/K_A$, but is also very accurate in assessing the variation of these estimations.

Mindell ([42]) conducted a study to assess evidence for positive selection in human and chimpanzee primate immunodeficiency viruses type 1 (PIV1s) by using ratios of synonymous to nonsynonymous nucleotide change. Evidence of the selective forces acting on a gene can be gathered by estimating the number of synonymous and nonsynonymous changes per site. If synonymous and nonsynonymous mutations are selectively equal, then the ratio $K_S/K_A$ is expected to be close to unity; but if purifying selection is preferentially eliminating amino acid changes, then the $K_S/K_A$ ratio will be significantly higher. A higher rate of nonsynonymous substitutions compared with synonymous (\( K_S/K_A < 1 \)) is evident of positive selection for amino acid change. In that study, estimations were obtained by applying MEGA, a computer package written by Kumar et al. ([33]). The mean values of $K_S/K_A$ ratio for pol, gag, env gp41, and env gp120 were computed to be 7.28, 5.47, 3.40, and 2.81, respectively, for all pairwise comparisons of the five most closely related human PIV1s (455, lai, csf, ndk, eli). Since negative (purifying) and positive selection are inversely related in that the former entails higher $K_S/K_A$ ratios with most nonsynonymous changes selected against, and the latter entails lower ratios with more nonsynonymous changes allowed or selected for, it was concluded that the decreasing $K_S/K_A$ ratios indicate the purifying selection is decreasing and positive selection is increasing in considering the genes in the order listed above. However, the variance of $K_S/K_A$ was not estimated by the author, therefore, the comparisons were largely intuitive.
Here we recompute the $K_S/K_A$ ratios and their corresponding estimated standard deviations for all pairwise comparisons for regions gag and pol of the genes listed above. We will be able to test two hypothesis: (1) If the $K_S/K_A$ ratio of each region of each pairwise comparison is significantly higher than 1; (2) If the $K_S/K_A$ ratio of different region of each pairwise comparison is decreasing as indicated by the mean $K_S/K_A$ ratios from the original paper. Results of our calculation for gag and pol genes are provided in Tables 3.10 and 3.11. According to the new method, the average estimated $K_S/K_A$ ratios for gag and pol are 3.86 and 5.19, respectively. Both are about 30% lower than the original estimates obtained by applying MEGA. This actually should not be a surprise. As Ina ([28]) pointed out, Nei and Gojobori's ([44]) method, which is employed in MEGA, always overestimates $K_S$ and underestimates $K_A$, thereby overestimates the ratio of $K_S/K_A$. The extent of the systematic biases in

<table>
<thead>
<tr>
<th></th>
<th>lai</th>
<th>csf</th>
<th>ndk</th>
<th>eli</th>
</tr>
</thead>
<tbody>
<tr>
<td>455</td>
<td>3.558±0.858 (2.22, 5.71)</td>
<td>4.332±1.028 (2.72, 6.70)</td>
<td>3.732±0.877 (2.35, 5.92)</td>
<td>4.231±1.016 (2.64, 6.77)</td>
</tr>
<tr>
<td>lai</td>
<td>5.305±2.197 (2.36, 11.94)</td>
<td>3.287±0.963 (1.85, 5.84)</td>
<td>3.413±1.063 (1.85, 6.29)</td>
<td></td>
</tr>
<tr>
<td>csf</td>
<td>3.884±1.147 (2.18, 6.93)</td>
<td>4.978±1.537 (2.72, 9.12)</td>
<td></td>
<td></td>
</tr>
<tr>
<td>ndk</td>
<td>1.886±0.748 (0.87, 4.10)</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Table 3.10: $K_S/K_A$ ratio for gag, the estimated ratio, the SD of the estimator, the suggested 95% confidence intervals for the parameter using the log transform, the mean ratios obtained by MEGA is 5.47.
Table 3.11: $K_S/K_A$ ratio for pol, the estimated ratio, the SD of the estimator, the suggested 95% confidence intervals for the parameter using the log transform, the mean ratios obtained by MEGA is 7.28.

<table>
<thead>
<tr>
<th>c/r</th>
<th>lai</th>
<th>csf</th>
<th>ndk</th>
<th>eli</th>
</tr>
</thead>
<tbody>
<tr>
<td>455</td>
<td>6.665 ±0.992 (4.98, 8.92)</td>
<td>6.149 ±0.896 (4.62, 8.18)</td>
<td>7.587 ±1.139 (5.65, 10.18)</td>
<td>6.294 ±0.911 (4.74, 8.36)</td>
</tr>
<tr>
<td>lai</td>
<td>3.098 ±0.850 (1.81, 5.31)</td>
<td>5.607 ±1.123 (3.79, 8.30)</td>
<td>4.659 ±0.890 (3.20, 6.78)</td>
<td></td>
</tr>
<tr>
<td>csf</td>
<td>4.859 ±0.912 (3.36, 7.02)</td>
<td>4.184 ±0.730 (2.97, 5.89)</td>
<td></td>
<td></td>
</tr>
<tr>
<td>ndk</td>
<td>2.792 ±0.726 (1.68, 4.65)</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

The estimation of $K_S$ and $K_A$, however, is not completely clear. The overestimation of $K_S$, according to Ina ([28]), ranges from about 16% to 38%, while the underestimation of $K_A$ is from 6% to 30%.

Tables 3.10 and 3.11 clearly show that the ratios of $K_S/K_A$ between all pairs, except for the gag gene between eli and ndk, are significantly larger than 1, which provides evidence of purifying selection for amino acid change between the majority of the pairs of gag, and all pairs of pol. Further comparison is made to determine if the difference of $K_S/K_A$ between gag and pol is due to chance. The pairwise comparisons between gag and pol regions

$$
\frac{\hat{K}_S/\hat{K}_A(gag) - \hat{K}_S/\hat{K}_A(pol)}{\sqrt{\text{var}(\hat{K}_S/\hat{K}_A(gag)) + \text{var}(\hat{K}_S/\hat{K}_A(pol))}}
$$
should be approximately standard normally distributed under the null hypothesis that the ratio of $K_S/K_A$ remains the same across the two regions of the genes. The results of this comparison are listed in Table 3.12. According to the study, only two pairs of pol genes, 455 and lai, 455 and ndk, are concluded to have a higher $K_S/K_A$ ratios than their gag genes counterparts. While the difference between pol and gag among other pairs can be explained by chance.

<table>
<thead>
<tr>
<th></th>
<th>lai</th>
<th>csf</th>
<th>ndk</th>
<th>eli</th>
</tr>
</thead>
<tbody>
<tr>
<td>455</td>
<td>-2.369*</td>
<td>-1.332</td>
<td>-2.682*</td>
<td>-1.512</td>
</tr>
<tr>
<td>lai</td>
<td>0.937</td>
<td>-1.568</td>
<td>-0.899</td>
<td></td>
</tr>
<tr>
<td>csf</td>
<td>-0.546</td>
<td>0.467</td>
<td></td>
<td></td>
</tr>
<tr>
<td>ndk</td>
<td></td>
<td>-0.869</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Table 3.12: Standardized difference between gag and pol genes, * indicates the difference is significant at $\alpha = 0.05$ level.
CHAPTER 4

Review of Molecular Phylogeny

Molecular phylogeny is the study of evolutionary relationships among organisms by techniques of molecular biology. DNA sequences have been used to resolve many of the long-standing problems in phylogenetic studies. The purpose of this chapter is to explain how to reconstruct a phylogenetic tree from molecular data in the framework of a statistical model, and to introduce a phylogenetic issue that will be addressed statistically later.

4.1 Maximum Likelihood Method

If one adopts a statistical point of view, the estimation of phylogenies can then be seen simply as making an estimation of an unknown quantity, in the presence of uncertainty. There exist numerous methods for inferring phylogenies from molecular data. Among them, the parsimony methods, the distance matrix methods, and the maximum likelihood methods are the most commonly used. Felsenstein [11] assessed the reliability of these methods. They were considered as methods of statistical inference and their statistical properties were investigated. The main property studied was the consistency. A statistical estimation method is consistent if it approaches the true value of the quantity as larger and larger amounts of data are accumulated.
Maximum likelihood methods are found to be consistent, with the exception of certain cases when the number of quantities being estimated rises at least at the same rate as the number of data points. Distance matrix methods are also found to be consistent if the distances are transformed so that their expected values are equal to the total branch lengths intervening between two species. In contrast, parsimony methods were found to be inconsistent even with a perfect molecular clock for five species by Hendy & Penny [23]. Nothing in general is known about the conditions for consistency of parsimony methods.

The maximum likelihood method is a general method of deriving statistical estimates. In essence it is quite simple — one has an evolutionary model (M) and data (D), the likelihood of a tree (T) is the probability of the data given the tree and the model, denoted as $P(D:T,M)$. It is usually considered as a function of the tree. The probability of all possible sets of data must add up to one, but when the data is held constant and the tree is varied, the different values of $P(D:T,M)$ need not add up to one and are called likelihoods rather than probabilities. The maximum likelihood method involves finding the tree which yields the highest likelihood of producing the observed data.

It was inevitable that the maximum likelihood method would be applied to phylogenetic studies. However, the maximum likelihood method was not used frequently because of computational difficulty. Felsenstein [10] developed a computationally feasible method for finding such maximum likelihood estimates. The method addressed the problem of inferring evolutionary trees from DNA sequences under a simple probabilistic model of DNA evolution. For the particular case in which rates of base substitution are allowed to differ among lineages, an iterative method of altering the
tree was also developed which guarantees a continued increasing in the likelihood.

First, a method for evaluating the likelihood of a given tree was developed. To facilitate the computation, it will be assumed that changes at different sites in the sequence are independent probabilistic events. By making this assumption, the probability of a given set of data arising on a given tree could be computed site by site, and the product of the probabilities will be taken across sites at the end of the computation. This is a restrictive assumption. In particular, deletion and insertion events, which usually involve adjacent sites, can not be adequately modeled by assuming independence of events at different sites. However, practical computation does not appear feasible without this assumption.

The computing of the likelihood can be best illustrated by looking at an example (see Figure 4.1). In practice, only the states (bases) at a particular site at points 1, 2, 3, 4, and 5 are observable, while at points 0, 6, 7, and 8 they are unknown. $s_i$ is

Figure 4.1: The tree used in the discussion of computing the likelihood. The V's are the lengths of the segments.
used to denote the state at point \( i \) on the tree. The likelihood of this given rooted tree is estimated as:

\[
L = \sum_{s_0} \sum_{s_5} P_{s_0,s_5}(v_0)[P_{s_5,s_1}(v_1)][P_{s_5,s_2}(v_2)] \sum_{s_7} P_{s_5,s_7}(v_3)[\sum_{s_7} P_{s_7,s_4}(v_4)(P_{s_7,s_5}(v_5))].
\] (4.1)

The \( \pi \)'s must be the prior probabilities of finding each of the four bases at point 0 on the tree. They reflect the overall base composition in the group under study. The parentheses in equation (4.1) bear an exact relationship to the topology of the tree, since it is \( \Pi(\Pi) \) there is one \( P \) for each segment of the tree.

A key result presented by Felsenstein ([10]) is the Pulley Principle: an unrooted tree has the same likelihood as an equivalence class of rooted trees, namely all those compatible with the given unrooted tree (different placement of a root to the unrooted tree yields a member of this equivalence class of rooted trees). This principle allows one to evaluate an unrooted tree instead of a whole class of the compatible rooted trees. However, since the number of rooted trees with \( n \) tips is the same as the number of unrooted trees with \( n + 1 \) tips, there are still an astronomical number of possible (unrooted) topologies for even a moderate numbers of OTUs. The strategy Felsenstein [10] used was to build the tree up by successively adding species to it, starting with a two-species tree. When the \( k \)th species was being added to the tree, there would be \( 2k - 5 \) segments from which it could arise. Each of these was tried and the maximum likelihood within the resulting topology was evaluated. The placement yielding the highest likelihood was accepted.

The approach proposed by Felsenstein [10] has the advantage of finding the "best" tree faster compared to previous maximum likelihood algorithms, and it is still valid even when the rates of evolution differ in different lineages. However, the strategy
is not guaranteed to find the best topology. In fact, it is only guaranteed that the likelihood is increased and has arrived at a stationary point. It could be a saddle-point rather than maximum. The result depends on the order in which the species are added to the tree, so that if it is repeated with a different ordering of the species, a different result may be obtained, especially with less self-consistent data.

Fukami-Kobayashi and Tateno [14] conducted computer simulations to evaluate the robustness of the maximum likelihood method in the estimation of molecular trees against different nucleotide substitution patterns. Two factors of interest were the transition versus transversion ratio (s-v ratio) and the base contents, in particular the GC content.

A computer program used for the simulation was developed in which the values of the two factors could be given at will. First, two model genes of 300 and 1500 base sites were created to see the effect of gene size on the performance of the maximum likelihood method. In both genes, the rates of base substitutions were kept the same. Three simulation tests were conducted:

1. transition versus transversion ratios were different while the base contents were the same;

2. base contents were different while transition versus transversion ratios were the same;

3. both transition versus transversion ratios and base contents were different.

Comparison between the cases of 300 and 1500 sites indicated that topology estimation was improved with an enlarging gene size. As to the topology estimation, the maximum likelihood method was shown to be considerably robust against varying
the $s$-$v$ ratio and GC content, unless these two factors deviated greatly from the true values. The branch lengths leading to the largest likelihood value for a given tree topology, however, were quite sensitive to both $s$ and GC content.

To evaluate the extent to which the maximum likelihood tree is a significantly better representation of the true tree than the other candidates, it is important to estimate the variance of the difference between the log likelihoods of different tree topologies. Bootstrap resampling can be used for this purpose, but it imposes a great computational burden. To overcome this difficulty, Hirohisa and Hasegawa [25] developed a method for estimating the variance by expressing it explicitly. This method, according to the authors, should give essentially the same result as the bootstrap.

Homologous sites are regarded as units of sampling. As a first approximation, sites are assumed independent of one another along the sequence. Suppose there are $s$ homologous sequences with $n$ nucleotide sites. The Kullback-Leibler information quantity

$$I(g; f) = \sum_{i=1}^{m} g(x_i) \log \left( \frac{g(x_i)}{f(x_i|\theta)} \right)$$

(4.2)

is used to measure the distance of $g$ from $f$, where $m = 4^s$, and the vector $\theta$ denotes the unknown parameters such as the branching dates and the base substitution rates along the respective branches of a tree. Smaller values of the Kullback-Leibler information quantity indicate that $f$ is close to $g$. The maximum likelihood estimator of a given tree topology $i$, $\hat{\theta}_i$, is a consistent estimator of $\theta^*_i$, the parameter that minimizes the Kullback-Leibler information quantity under a weak condition. Therefore, the authors argued that even when different tree topologies do not constitute nested models, based on the Kullback-Leibler information, the maximum likelihood is still a good measure for comparing different topologies.

65
It is usually more convenient to work with the natural logarithm of the likelihood in molecular studies. For a large sample, the maximum log likelihood of a given tree topology \(i\)

\[
l_i(\hat{\theta}_i|\mathbf{X}) = \sum_{j=1}^{n} \ln f_i(X_j|\hat{\theta}_i) \tag{4.3}
\]

follows the normal distribution asymptotically from the central limit theorem, where \(\mathbf{X}\) is the whole data and \(X_j\) is the value of \(j\)th site. Its variance is estimated from the sample variance by

\[
\hat{\text{Var}}[l_i(\hat{\theta}_i|\mathbf{X})] = \frac{n}{n-1} \sum_{j=1}^{n} [\ln f_i(X_j|\hat{\theta}_i) - \frac{1}{n} \sum_{k=1}^{n} \ln f_i(X_k|\hat{\theta})]^2. \tag{4.4}
\]

In the same way, the difference of the maximum likelihood \(l_2(\hat{\theta}_2|\mathbf{X}) - l_1(\hat{\theta}_1|\mathbf{X})\) follows the normal distribution asymptotically, and its variance is estimated by

\[
\hat{\text{Var}}[l_2(\hat{\theta}_2|\mathbf{X}) - l_1(\hat{\theta}_1|\mathbf{X})] = \frac{n}{n-1} \sum_{j=1}^{n} [\frac{f_2(X_j|\hat{\theta}_2)}{f_1(X_j|\hat{\theta}_1)} - \frac{1}{n} \sum_{k=1}^{n} \frac{f_2(X_k|\hat{\theta}_2)}{f_1(X_k|\hat{\theta}_1)}]^2. \tag{4.5}
\]

One use of the maximum likelihood method is to test whether particular phylogenetic models provide a good explanation of observed sequences. The test can be extended to evaluate specific components of models, for example, the assumption of a molecular clock. In traditional statistical theory, a widely accepted statistic for testing the goodness of fit of models is the likelihood ratio statistic \(2\Delta = 2\ln(\hat{L}_1/\hat{L}_0) = 2(\hat{S}_1 - \hat{S}_0)\), with \(S\) being the natural logarithm of the likelihood \(L\). When \(H_0 \subset H_1\) and the null hypothesis is correct, this statistic is asymptotically distributed as \(\chi^2_k\), irrespective of the true value of any parameters of the null hypothesis. The number of degrees of freedom (d.f.), \(k\), is the difference between the number of free (estimated) parameters in \(H_1\) and \(H_0\). However, a \(\chi^2\) distribution cannot be justified in the
present phylogenetic context for two reasons, namely, the conditions to ensure the \( \chi^2 \) distribution are not satisfied, and the approximate degrees of freedom, \( k \), is not clear in tree-inference problems. More insight can be found in Goldman [16].

One of the statistical tests based on the maximum likelihood method was derived by Cox [6] [7]. Cox considered the case in which the hypotheses are not in the nested form \( H_0 \subset H_1 \) and the \( \chi^2 \) approximation for the null hypothesis distribution of \( 2\Delta \) does not apply. Cox’s statistic is a likelihood ratio statistic in which the likelihood is maximized separately under each of the two hypothesis, and the ratio of likelihoods is formed:

\[
\delta = \ln \left[ \frac{\sup_{\beta \in H_1} L_1(x, \beta)}{\sup_{\alpha \in H_0} L_0(x, \alpha)} \right] = S_1(x, \hat{\beta}) - S_0(x, \hat{\alpha}).
\]

It is crucially important that these likelihoods be maximized independently for each \( i \) when performing the Monte Carlo simulation. In other words, the form

\[
\delta_i = \ln \left[ \frac{\sup_{\beta \in H_1} L_1(x_i, \beta)}{\sup_{\alpha \in H_0} L_0(x_i, \alpha)} \right]
\]

must be used, and not \( \delta_i = \ln [L_1(x_i, \hat{\beta})/L_0(x_i, \hat{\alpha})] \) (Hall and Wilson [18]).

4.2 Methods for Detecting Recombination

Human immunodeficiency viruses (HIVs) exhibit extraordinary genetic diversity. This is, in part, due to recombination (Robertson et al. [47]). One way to detect recombination is in the context of phylogenetic analysis. However, early methods for detecting recombination among DNA sequences were largely intuitive, and it is not always clear whether unusual similarity between two DNA sequences is due to chance, gene conversion, or functional restraint. Some rigorous statistical methods were developed later (Stephens [58]; Sawyer [52]; Smith[56]), based on mapping putative
crossover points of recombination events by examining the distribution of phylogenetically informative sites supporting alternative tree topologies.

Stephens [58] has developed statistical tests that aid in deciding whether or not particular sets of nucleotide sites are statistically clustered and use this information to make qualitative inferences about recombination events. The tests are based on the linear distribution of a set of $s$ specific sites which share a common characteristic among a total of $n$ sites. When the tests are applied to detect gene conversion and intragenic recombination in a sample of DNA sequences, the sites of interest are those that correspond to the particular ways of splitting the sequences into exactly two groups (e.g., sequences A and D vs sequences B, C, and E-J). Each such split is termed a phylogenetic partition. Variable nucleotide sites usually will create phylogenetic partitions in the sense that all sequences that share one variant at a particular site are distinguished from those sequences that share an alternative nucleotide. All sites creating the same partition are called congruent sites and are referred to as being congruent to the partition as well as to each other. A partition is call primary if it corresponds to unique, undisturbed mutations, as site $j$ in Figure 4.2, whereas a secondary partition would be one created by multiple events occurring at a given site, e.g., site $i$ in Figure 4.2.

Stephens [58] declared that the clustering of sites that are congruent to secondary partitions indicates that such partitions were created by recombination events. The basic idea of the test is that clustering of a set of $s$ variable sites may be tested under the assumption that such sites are randomly distributed along the DNA sequences. Let $d$ be the length of the interval spanned by all $s$ variable sites. The probability that $d$ is less than or equal to the observed distance $d_0$, which, according to Stephens
Figure 4.2: Effect of different mutational events on labeling of partitions. A-H are sequences; i and j are nucleotide sites. Parallel mutations occur at site i; unique mutation occurs at site j. Partitions: i=ABD/CEFGH; j=ABCDE/FGH.

\[ P(d \leq d_0) = \left[ s - (s - 1)(d_0 + 1)/n \right] \prod_{j=0}^{s-2} \frac{d_0 - j}{n - j - 1}. \] (4.6)

It is useful for testing whether or not a set of s specified sites is clustered relative to the regions flanking them. However, this test may fail to show clustering even when the clustering of some of the sites is apparent because the group of s-sites is bounded on either or both sides by many consecutive non-s sites. In other words, one may face the question whether it is the runs of s-sites or the runs of non-s sites (or both) that are too long. One approach to resolve this dilemma is to test runs of the s sites after first excluding the most unusual long non-s segments. The probability that at least one of the s - 1 random, independently observed unvaried segments would be as long
or longer than the longest observed segment $g_0$, approximately, can be computed as

$$ P \approx (s - 1) \sum_{k=0}^{r} g_k $$

where $g_k$ is the probability that exactly $k$ of the $r$ unvaried sites between a randomly chosen pair of consecutive $s$ sites. It is evaluated as

$$ g_k = \binom{s + r - k - 3}{r - k} / \binom{s + r - 2}{r} $$

Equation (4.7) can then be used to test whether the segments of consecutive non-$s$ sites bounding the group of $s$-sites are unusually long. If a segment of non-$s$ sites is judged by (4.7) to be too long, then it will be excluded, with the closest segment of $s$-sites, from further analysis and the probability of observing a segment of non-$s$ sites as long as or longer than the next longest segment observed will be recalculated. After all the unusually long non-$s$ segments are excluded, equation (4.6) can be used to test whether or not a set of remaining $s$ specific sites is clustered relative to the region flanking them. A detected strong clustering of these specific sites that are congruent to secondary partitions may indicate that recombinational events have occurred. However, if the sample has more than three or four sequences, and if the sequences are moderately or highly polymorphic, an appropriate phylogenetic partition which will split the species into exactly two groups may be hard to find. In practice, it is usually difficult to know which phylogenetic partitions are primary and which are secondary, especially when dealing with intraspecific DNA sequences. Also, phylogenetic partitions are treated individually and statistical concerns about multiple comparisons may arise. For example, if there are $m = 8$ sequences, then there are $2^{m-1} - 1 = 127$ possible phylogenetic partitions that will split the sequences into two subsets, and some of these partitions may have significantly nonuniform
distributions purely by chance.

Sawyer [52] also developed a test which is based on imbalances in the distribution of segments on which some pair of sequences agree. The essence of this method is to consider polymorphic sites, those sites known to vary in at least one of a set of strains, and to ask whether the differences between pairs of strains tend to occur in runs. If two (out of the }\textit{n}\text{) sequences are compared, they will differ in a set of }d < s\text{ polymorphic sites, where }s\text{ is the total number of polymorphic sites. These }d\text{ discordant sites partition the genome into }d+1\text{ subsets, which are called the fragments determined by this partition. A condensed fragment is the set of polymorphic sites in a fragment. Hence the length of a condensed fragment is the number of polymorphic sites either between neighboring discordant sites or between a discordant site and one of the ends of the sequence. The sum of the lengths of the condensed fragments is }\sum x_i = s - d,\text{ where }x_i\text{ is the length of the }i\text{th condensed fragment. Two statistics were proposed to detect the recombination events:}

- MCF : the maximum of }x_i\text{ for all condensed fragments for all pairs of sequences;}
- SSCF : sum of the squares of the condensed fragment lengths, which is defined as the sum of }x_i^2\text{ over all }d_k + 1\text{ fragments over all }n(n - 1)/2\text{ pairs of sequences, where }d_k\text{ is the number of discordant sites determined by the }k\text{th pair of sequences.}

To estimate the significance of MCF and SSCF, critical values are obtained through simulation method. Any base position that is not identical in all }n\text{ DNA sequences and that is not part of a codon position that encodes different amino acids is called a silent polymorphic site. The scores for the observed sequences are compared with
those for artificial data sets obtained from many random permutations of the order of the \( s \) silent polymorphic sites. The permutation is created in such a way that for each pair of sequences, the number of discordant sites \( d_k \) and the sum of the fragment lengths \( \sum x_i = s - d_k \) is preserved, although \( \sum x_i^2 \) will generally vary. \( P \)-values for SSCF and MCF are defined as the proportion of permuted data sets that have SSCF or MCF scores, respectively, that are greater than or equal to the original score. Significant small \( P \)-values will lead to the conclusion of recombination. Not surprisingly, SSCF is found to be a more powerful statistic for testing the recombinational events. Since the relationship of the recombinational events and the distribution of those segments considered is not completely clear, extra work is needed to investigate the biological reasoning behind this method.

To demonstrate mosaic structure statistically and determine the limits of the blocks, Smith [56] proposed the maximum chi-squared method. This method is mostly useful when two parental sequences and a derived sequence are available. A region of DNA containing \( n \) polymorphic sites is considered. First a pair of sequences, which differ at \( s \) sites are compared. If an arbitrary cut after the \( k \)th site is made, then the sequences differ at \( r \), say, of the \( k \) sites before the cut, and \( (s - r) \) of the \( (n - k) \) sites after the cut. The value, \( c \), of the \( 2 \times 2 \chi^2 \) statistic is then calculated. If the difference between two sequences has a mosaic structure, with only two blocks, then the optimal boundary between the blocks occurs after site \( k_{\text{max}} \), for which the value of \( c \) is a maximum, \( c_{\text{max}} \). To test whether this division is significant, Smith [56] suggested constructing \( T \) sequence pairs of length \( n \), different at \( s \) randomly distributed sites, and for each pair find \( c_{\text{max}} \). If the value of \( c_{\text{max}} \) for the real data is greater than the value for any trial pair, then the observed mosaic structure is said to be significant at
the level \( P < 1/T \). In determining the block of one sequence, it will be compared to all the other sequences under study. However, the comparisons need not give the same answer, namely, some may conclude that the division is significant, while others may not. And even if all the comparisons are significant, the cuts identified may still be in different positions. This procedure works in practice only if the appropriate structure consists of just two blocks, with one single division. Although, in principle, it could be extended to two or more divisions, the computation time would be prohibitive.

Hein [21] also developed an explicit algorithm for detecting recombination. The algorithm finds a set of trees describing the history of a given set of sequences. Recombinations are concluded to have occurred if the history cannot be explained by a single phylogeny. The method is a generalization of the parsimony method that is used to analyze sequences without recombination.

A set of \( n \) sequences \( S_1, \ldots, S_n \) of length \( l \) is assumed given. Positions that partition the sequences into groups, where at least two groups consist of more than one sequence, are frequently called informative positions. In this analysis only informative positions are considered. The overall strategy in constructing the algorithm is to convert the problem into a shortest-path problem in a graph (see Figure 4.3).

For each informative position, create as many nodes as there are for potential trees relating the sequences. Let \((i, t)\) be the node at position \( i \) (ith column in the alignment) corresponding to a topology \( t \). This node is assigned weight \( w(i, t) \), the substitutional weight of position \( i \) given topology \( t \). An algorithm calculating \( w(i, t) \) was worked out by Fitch [12], Hartigan [19], and Sankoff [51] (abbreviated the FHS-algorithm). Two neighboring nodes \((i \text{ and } i - 1)\) are connected by an edge and are assigned weight \( d(t, t') \), the recombinational distance between \( t \) and \( t' \). It is weighted
Figure 4.3: Shows the dynamic programming argument. Each node represents a topology; all nodes at the \((i - 1)\)st column are connected with all nodes at the \(i\)th column.

Proportionally to the number of recombinations needed to transform the two topologies connected by the edge. \(W(i, t)\) is the weight of the most parsimonious history of the first \(i\) positions given that the topology of position \(i\) is \(t\). The \(W(i, t)\)s will obey the following recursion:

\[
W(i, t) = \begin{cases} 
  \min\{W(i - 1, t') + d(t, t') + w(i, t)|t'\} & \text{if } i > 1 \\
  w(1, t) & \text{if } i = 1
\end{cases}
\]

The most parsimonious history of sequences will then correspond to the shortest path starting at the first informative position and ending in the last. The weight of a path is the sum of the weight of nodes and edges on it.

However, this method becomes impractical for more than five or six sequences due to the large number of topologies. To overcome this difficulty, Hein [22] proposed a heuristic algorithm that can analyze data consisting of more than six sequences. The main assumption is that recombinations happen at most once between every nucleotide, hence it is only necessary to investigate neighbor trees that are one recombination apart. First the phylogeny for the complete sequences ignoring the
problem of recombination is obtained. This phylogeny is then used as the tree for
the leftmost part. The approximate algorithm is then used, scanning the sequences
from left to right, continuously investigating trees that are one recombination away
from the given tree. If one of these trees is better, then it is chosen for the region,
and then trees in the neighborhood of this tree are investigated, etc. Since the choice
of the starting phylogeny is partly arbitrary, this algorithm is run first from the left
to right and then from right to left. The assumption behind this is that the tree
found by the algorithm for the rightmost point is a better starting phylogeny than
the original one. This algorithm will find a series of local phylogenies, subject to
some constraints. That the chosen local phylogenies are better than their neighbor
phylogenies is illustrated by checking their weight for the regions.

Reconstructing phylogenies by parsimony has been severely criticized (Felsenstein
[11]). Its main flaws are: (1) severe underestimation of branch lengths; (2) inconsis-
tency - in many cases the method will converge toward a wrong tree as data accumu-
late; (3) lastly, no estimate of reliability is given. All these problems are undoubtably
retained when generalizing the method to include recombinations. The robustness of
Hein’s [22] algorithm needs to be examined too, i.e., how the choice of the starting
tree will affect the outcome of the final result is a very important question needing
to be answered. When the hypothesis of whether there is a recombination is tested,
the comparison of weights between two topologies for the same region is done from
the viewpoint of parsimony but not of statistical inference. Even if different trees
are found to describe the history of the sequences, is it because of recombination or
purely due to chance? This method provides no answer to this question.
There is a more general reason for doubting whether Hein's [22] method is appropriate. Suppose that two sequences, A and B, differ by say, 15% of nucleotides, and that C is identical to A in the upstream region and to B in the downstream region. This is unequivocal evidence for recombination, but Hein's [21], [22] method is unable to recognize this simplest type of evidence for recombination since there are no informative sites. If the phylogenetic method is to be used, therefore, it needs to be modified to allow not only for informative sites, but for at least all polymorphic sites.

Templeton et al. [62] presented a strategy that uses the network concept instead of trees for estimating the set of cladograms that are consistent with a particular sample of either restriction site or nucleotide sequence data. This estimation procedure also identifies haplotypes that are candidates for being products of recombination.
CHAPTER 5

Detecting Discordant Phylogenetic Trees

There are many methods to detect recombination events. We approach this problem as a standard statistical testing problem, with the null hypothesis being that the phylogenetic relationships among the two sets of sequences are the same, while the alternative is that they are different. The test can be done by comparing the inferred phylogenetic trees from the two sets of the sequences. We will employ the maximum likelihood method for this comparison. The test statistic is constructed in a manner to ensure the independence of the nucleotides compared and the inferred tree on which the comparison is drawn.

The theory we will develop in this chapter is based on a large sample approximation. For our case, larger samples mean longer sequences, but not necessarily more species. We assume that nucleotide sites develop independently of one another along the sequences. Homologous sites can then be regarded as units of sampling. It is unlikely that any sequences will fully satisfy this assumption, but non-independent models are currently intractable. Before formally introducing our approach, some notation is needed. Let $D_1$, $D_2$ be two sets of $N$ homologous aligned DNA sequences of
the population of interest drawn from two different sources. A procedure for discrimination between the phylogenetic relationships among the two sources, from which $D_1$ and $D_2$ are drawn, is derived. The data $D_1$ and $D_2$ can be represented as follows:

$$
\begin{array}{cccccccc}
\text{species 1:} & X_{11} & X_{12} & \ldots & X_{1u} & Y_{11} & Y_{12} & \ldots & Y_{1v} \\
\text{species 2:} & X_{21} & X_{22} & \ldots & X_{2u} & Y_{21} & Y_{22} & \ldots & Y_{2v} \\
\vdots & \vdots & \vdots & & \vdots & \vdots & \vdots & & \vdots \\
\text{species N:} & X_{N1} & X_{N2} & \ldots & X_{Nu} & Y_{N1} & Y_{N2} & \ldots & Y_{Nv}
\end{array}
$$

where the X's or Y's are either A, G, C, or T. We write the ith site $(X_{i1}, \ldots, X_{Ni})^T$ $[(Y_{i1}, \ldots, Y_{Ni})^T]$ (a superscript T denotes a transpose vector) as $X_i [Y_i]$. Then the likelihood for either data set, say $D_1$, can be written as

$$
L(\theta|X) = \prod_{i=1}^{u} f(X_i|\theta)
$$

where $f(X_i|\theta) = f(x_1, x_2, \ldots, x_N|\theta)$ is the probability that the species 1 has the base $x_1$, the species 2 has $x_2$, ..., and the species $N$ has $x_N$ at a homologous site. The vector $\theta$ denotes unknown parameters such as the tree topology and the branch lengths.

We assume the Markov property for the process of base substitutions in our model. The transition probability matrix $P(t)$ is represented by $P(t) = e^{tR}$, where the matrix $R$ consists of the instantaneous rates of substitution, $r_{ij}$, from the base $i$ to the base $j$. The diagonal element $r_{ii}$ is defined so that the sum of each row is equal to zero.

Felsenstein uses the following model for $R$ in his program package PHYLIP (version 2.6 and later) which will be used in our approach:

$$
r_{ij} = \begin{cases} 
(k/\Pi_j + 1)u\pi_j & \text{for transition} \\
u\pi_j & \text{for transversion}
\end{cases}
$$

where $u$ is a parameter that determines transversion mutation rate, $k$ is a parameter that determines the transition/transversion rate ratio, therefore the transition mutation rate is implicitly determined. $\pi_j$ is the equilibrium proportion of base $j$ ($j=A,$
G, C, or T), and \( \Pi_j = \pi_T + \pi_C \) when \( j \) is a pyrimidine and \( \pi_A + \pi_G \) when \( j \) is a purine.

We shall employ the maximum likelihood method to detect if phylogenetic relationships for different parts of the genes are discordant which, according to Robertson et al. ([47]), serves as the most clear indication of the existence of a mosaic viral genome, i.e., one that has been generated by a recombination event at some point in the past. The relationships among the genes are actually unknown. However, they can be estimated by the derived evolutionary trees based on the genes. Therefore, we can go on to estimate phylogenies for different parts of the gene sequences, and search for discrepancies between these inferred trees. It should be pointed out that the ultimate goal is not to compare the estimated phylogenetic trees. Instead, it is to compare the relationships between different parts of the genes, which is approximately illustrated by the inferred trees based on different parts of the genes.

One crucial technique used in our proposed method is similar to that employed by the jackknife. Today, data-resampling methods are being used increasingly in statistical analysis. The resampling methods replaces theoretical derivations required in applying traditional methods by repeatedly resampling the original data and making inferences from the resamples. One of the most commonly used resampling techniques is the jackknife. Because of the availability of inexpensive and fast computing, this computer-intensive method has caught on very rapidly in recent years.

Quenouille [45] introduced a method, later named the jackknife, to estimate the bias of an estimator by deleting one datum each time from the original data set and recalculating the estimator based on the rest of the data.

Let \( T_n = T_n(X_1, \ldots, X_n) \) be an estimator of an unknown parameter \( \theta \), and \( T_{n-1}^i = T_{n-1}(X_1, \ldots, X_{i-1}, X_{i+1}, \ldots, X_n) \) be the given statistics based on \( n - 1 \) observations
The jackknife estimate of $\theta$ is
\[ T_{jack} = nT_n - (n - 1)\bar{T}_n, \]
where $\bar{T}_n = \frac{\sum T_{n-1}}{n}$. Suppose
\[ \text{bias}(T_n) = \frac{a}{n} + \frac{b}{n^2} + O\left(\frac{1}{n^3}\right) \]
where $a$ and $b$ are unknown but do not depend on $n$, then
\[ \text{bias}(\bar{T}_n) = \frac{a}{n} + \frac{b}{(n - 1)(n - 1)^2} + O\left(\frac{1}{(n - 1)^3}\right), \]
\[ \text{bias}(T_{jack}) = -\frac{b}{n(n - 1)} + O\left(\frac{1}{n^3}\right), \]
i.e., the bias of $\bar{T}_n$ is equivalent to that of $T_n$, the bias of $T_{jack}$ is of order of $n^{-2}$. In other words, $E(\bar{T}_n) = E(T_n)$, and the jackknife produces a method to eliminate a bias term of $O\left(\frac{1}{n}\right)$ in the problem with $n$ observations.

The variance of an estimator is sometimes hard to obtain, especially when the estimator and the underlying distribution are complex. The jackknife has become a more valuable tool since Tukey [65] found that the jackknife can also be used to construct variance estimators. Tukey defined
\[ \tilde{T}_n^i = nT_n - (n - 1)T_{n-1}^i, i = 1, \ldots, n \]
as the jackknife pseudovalues and conjectured that
(A) The pseudovalues $\tilde{T}_n^i, i = 1, \ldots, n$, may be treated as though they were i.i.d.;
(B) $\tilde{T}_n^i$ has approximately the same variance as $\sqrt{n}T_n$.

Hence it is natural to estimate $\text{var}(T_n)$ by
\[ v_{jack} = \frac{n - 1}{n} \sum_{i=1}^{n} (T_{n-1}^i - \frac{1}{n} \sum_{j=1}^{n} T_{n-1}^j)^2. \]
This estimator is the well-known jackknife variance estimator for $T_n$. There are many studies and applications of the jackknife. Miller [40] provided rigorous justification for the use of the jackknife to obtain an approximate $t$-statistic to be used in testing or for the construction of confidence intervals. For the case where $T_n$ is a function of sample means, or a function of a maximum likelihood estimate, or a function of regression estimates, $(T_{\text{jack}} - \theta)/\sqrt{v_{\text{jack}}}$ is shown to have an approximate $t$ distribution, or for large $n$, an approximate normal distribution by Miller [41]. Thorburn [63] also investigated the conditions under which the jackknife estimator is asymptotically normal distributed, and the variance estimator based on the pseudovalues is consistent. He also showed that if the jackknife and the original statistic have the same asymptotic distribution, it is often a normal one; while if the limiting distribution is not normal, then the jackknife and the original statistic differ significantly in most cases. In general, Shao and Tu [54] showed that $(T_n - \theta)/\sqrt{v_{\text{jack}}} \to N(0, 1)$ if $v_{\text{jack}}$ is weakly consistent: $v_{\text{jack}}/\text{var}(T_n) \overset{p}{\to} 1$, or strongly consistent: $v_{\text{jack}}/\text{var}(T_n) \overset{a.s.}{\to} 1$. To improve the accuracy of the normal approximation to the distribution of the studentised statistic, $(T_n - \theta)/\sqrt{v_{\text{jack}}} \to N(0, 1)$, Hinkley and Wei [24] and Tu and Zhang [64] suggested a method based on the estimation of the skewness of $T_n$ by using a weighted sum of delete-1 and delete-2 jackknife pseudovalues.

The jackknife is less dependent on model assumptions and does not need the theoretical formula required by the traditional approach. In terms of their asymptotic behaviors, the jackknife and the traditional estimators are usually first order asymptotically equivalent. Some limited empirical results show that the jackknife variance estimator is less biased but more variable than the traditional estimator. The jackknife variance estimator $v_{\text{jack}}$ is consistent for many statistics including functions of
the sample mean, U-statistics, M-estimators, some statistics based on generalized empirical distribution, and statistics that are continuous transformations of several statistics. The consistency of $v_{jack}$ usually requires more stringent smoothness conditions on the given statistic $T_n$ than the asymptotic normality of $T_n$ does. For its consistency, the delete-d jackknife variance estimator $v_{jack-d}$ requires less stringent smoothness conditions on the given statistic $T_n$ than $v_{jack}$ when $d > 1$. However, the computational complexity increases rapidly as $d$ increases. The asymptotic validity of the jackknife is justified for a smooth $T_n$ (see Shao and Tu [54] for more discussion).

5.1 The Likelihood-Based Hypothesis Test

We are interested in whether the underlying population form the same relationships at two different sources. This can be done, roughly speaking, by checking if the inferred phylogenetic trees have the same topology. A model of current interest will form the null hypothesis ($H_0$) and the alternative hypothesis ($H_1$) for the test:

$H_0$: two sets of sequences, whose sites have evolved independently, are related by the same (unknown) phylogenetic “tree” structure

$H_1$: two sets of sequences, whose sites have evolved independently, are related by different (unknown) phylogenetic “tree” structures

That the sites evolved independently will not be tested since it appears in both $H_0$ and $H_1$. It is, instead, an assumption we need to make to carry out the test.
5.1.1 Intuition

The procedure of statistical hypothesis testing usually consists of three steps:

1. set up the null hypothesis $H_0$, and alternative hypothesis $H_1$;

2. derive test statistics whose distribution under $H_0$ is known or can be approximated;

3. draw conclusions based on the critical value.

Since the hypotheses are already set up, our next step is then to find a statistic that can distinguish $H_0$ and $H_1$. One easy candidate for our test statistic is the difference between the average maximum likelihood values of the two sets of sequences for the two inferred trees. Since each site is independent by the assumption, this statistic plays a role similar to the difference between two sample means. Therefore under the null hypothesis, it should be expected to be close to zero, and the problem will, approximately speaking, become the standard two sample test. However, this comparison is meaningful only if the two inferred trees are identical. Otherwise, we may not only measure the difference between the two sets of nucleotide sequences, but also that between different tree topologies. In other words, the comparison based on Felsenstein’s likelihood is made conditioning on the inferred trees. However, there is no guarantee that identical inferred trees will be reconstructed if the two samples were drawn from the same phylogenetic tree. The reason is due to the huge number of possible inferred trees. The number of bifurcating rooted trees ($N_R$) for $n$ OTUs is given by

$$N_R = \frac{(2n - 3)!}{2^{n-2}(n - 2)!}$$
and the number of bifurcating unrooted trees ($N_U$) for $n \geq 3$ is

$$N_U = \frac{(2n - 5)!}{2^{n-3}(n - 3)!}.$$  

We can see that both $N_U$ and $N_R$ increase very rapidly with $n$. The numbers of possible rooted and unrooted trees for up to 10 OTUs are given in Table (5.1). Note that the number of possible unrooted trees for $n$ OTUs is equal to the number of possible rooted trees for $n - 1$ OTUs. Since only one of these trees is chosen as the inferred tree, it is usually very unlikely that the same tree will be chosen twice when $n$ is large.

The ideal situation will be that we can compare the maximum likelihood of

<table>
<thead>
<tr>
<th>Number of OTUs</th>
<th>Number of rooted trees</th>
<th>Number of unrooted trees</th>
</tr>
</thead>
<tbody>
<tr>
<td>2</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>3</td>
<td>3</td>
<td>1</td>
</tr>
<tr>
<td>4</td>
<td>15</td>
<td>3</td>
</tr>
<tr>
<td>5</td>
<td>105</td>
<td>15</td>
</tr>
<tr>
<td>6</td>
<td>954</td>
<td>105</td>
</tr>
<tr>
<td>7</td>
<td>10,395</td>
<td>954</td>
</tr>
<tr>
<td>8</td>
<td>135,135</td>
<td>10,395</td>
</tr>
<tr>
<td>9</td>
<td>2,027,025</td>
<td>135,135</td>
</tr>
<tr>
<td>10</td>
<td>34,459,425</td>
<td>2,027,025</td>
</tr>
</tbody>
</table>

Table 5.1: Possible numbers of rooted and unrooted trees for 2-10 OTUs.

each site of $D_1$ and $D_2$ based on the same topology and the difference of their average maximum likelihood follows a known distribution under the null hypothesis. By same topology, we mean the relationships among the OTUs are identical, but the branch
lengths can be different. We will propose an approach that can achieve this aim in the following section.

5.1.2 Theory

Some sites are of particular interest when detecting the discrepancy between the phylogenetic structures. We term a site a polymorphic site if not all sequences have the same nucleotide at that site. The polymorphic sites play a very important role in our approach. Among all the sites of $D_1$ and $D_2$, usually only a portion of them are polymorphic. Let us denote their numbers to be $m$ and $n$, respectively. We also use $T_1$, $T_2$ to represent the true phylogenetic tree relationships among the population of interest from source 1 and source 2, from which the two samples $D_1$ and $D_2$ are drawn, respectively. Two slightly different approaches to estimating these parameters are to use (1) $\hat{T}_1$, $\hat{T}_2$: the inferred trees constructed from $D_1$, $D_2$ respectively; or (2) $\hat{T}^1_i$: the inferred tree constructed from $D_1$ with a polymorphic site $X_i$ excluded. $\hat{T}^2_j$: the inferred tree constructed from $D_2$ with a polymorphic site $Y_j$ excluded. ($i = 1, \ldots, m; j = 1, \ldots, n$). The reason that we only consider the case with the polymorphic sites removed will be discussed in more detail later in this section.

One of the qualities that can be used to measure the difference between the two underlying phylogenetic relationships is the difference of the maximum likelihood values between the two sets of nucleotide sites evaluated based on the same topology. Conventionally, however, log likelihood values are usually used. Let us first define the two parameters of interest:

$$\gamma_1 = \frac{L(D_1|T_1) - L(D_2|T_1)}{N}$$

$$\gamma_2 = \frac{L(D_2|T_2) - L(D_1|T_2)}{N}$$

85
where $L(D|T)$ is used to represent the expected log likelihood per polymorphic site of set $D$ for the true but unknown phylogenetic relationship $T$ throughout this chapter. These two parameters measure the average difference between the two sources by taking the number of sequences into account. Obviously, under the null hypothesis that $D_1$ and $D_2$, have the same relationships among them, i.e., $T_1 = T_2$, then $\gamma_1 = \gamma_2 = 0$.

Because of the obvious systematic similarity between $\gamma_1$ and $\gamma_2$, we will focus on finding the estimator of $\gamma_1$ and its properties first. $\gamma_1$ is said to be estimable of degree $(1,1)$ since for any two polymorphic sites $X_\alpha$ and $Y_\beta$, $h_0(X_\alpha; Y_\beta) = \frac{l(X_\alpha|T_1) - l(Y_\beta|T_1)}{N}$ has the property that $E(h_0(X_\alpha; Y_\beta)) = \gamma_1$, where $l(X|T)$ is used to denoted the maximum log likelihood of $X$ for given topology $T$ throughout the chapter unless noticed otherwise. Then a two-sample U-statistic for estimating the parameter $\gamma_1$ can be constructed as

$$U_0(T_1, T_1) = \frac{1}{mn} \sum_{\alpha \in A} \sum_{\beta \in B} h_0(X_\alpha; Y_\beta)$$

(5.3)

where $A[B]$ is the collection of all single values chosen without replacement from the integers $1, \ldots, m[1, \ldots, n]$ (Randles and Wolfe [46]).

However, usually it is hard to evaluate (5.3) since the true tree relationship $T_1$ is normally unknown. To overcome this difficulty, replacing $T_1$ with its traditional estimator $\hat{T}_1$ in (5.3) is an easy choice to avoid the problem. This approach may be valid in some cases, but it creates another problem in general: the maximum log likelihood evaluated based on $\hat{T}_1$ is more likely to yield a higher value for the units of $D_1$, from which $\hat{T}_1$ is constructed, than for the units of $D_2$. Therefore the estimator is biased even under the null hypothesis. Under the null hypothesis, $\gamma_1$ is 0 while the U-statistic is very likely to have a positive value. The reason for this biased
estimation is the dependence between $T_1$ and $D_1$. To construct an inferred tree that is independent of the sample unit whose maximum log likelihood is computed, we employ a resampling technique that is similar to the jackknife. We first define a data set, say $D$, with a sample unit point, say $X_i$, removed the jackknife set (created by $X_i$).

To estimate the true phylogenetic tree $T_1$, we shall use the estimated phylogenetic tree $T_1^i$ based on the jackknife sets (created by $X_i$’s), $(i = 1, \ldots, m)$.

Instead of computing the U-statistic based on $h_0(X_\alpha; Y_\beta) = \frac{l(X_\alpha|T_1) - l(Y_\beta|T_1)}{N}$ or $h_0'(X_\alpha; Y_\beta) = \frac{l(X_\alpha|T_1^i) - l(Y_\beta|T_1^i)}{N}$, we define

$$h_1(X_\alpha; Y_\beta) = \frac{l(X_\alpha|T_1^\alpha) - l(Y_\beta|T_1^\alpha)}{N}.$$ 

Thus, the two-sample U-statistic becomes

$$U_1(\hat{T}_1^i, \hat{T}_1^i) = \frac{1}{m n} \sum_{\alpha \in A} \sum_{\beta \in B} h_1(X_\alpha; Y_\beta)$$

$$= \frac{1}{N} \frac{1}{m n} \sum_{\alpha \in A} \sum_{\beta \in B} (l(X_\alpha|T_1^\alpha) - l(Y_\beta|T_1^\alpha))$$

$$= \frac{1}{N} \frac{1}{m n} \sum_{i=1}^{m} \sum_{j=1}^{n} (l(X_i|T_1^i) - l(Y_j|T_1^i)). \quad (5.4)$$

In other words, the maximum log likelihood of each polymorphic site $X_i$ now is evaluated based on an inferred tree that is independent of $X_i$. Consequently, the maximum log likelihoods are not computed for a tree that favors one of the two samples. Though (5.4) is an unbiased estimator of $\gamma_1$, it is actually practically difficult to use because of the computational burden. To see this, we can express the double sum on the right hand side of (5.4) as follows:

$$RHS = l(X_1|T_1^1) + l(X_2|T_1^2) + \cdots + l(X_m|T_1^m)$$

$$- l(Y_1|T_1^1) - l(Y_1|T_1^2) - \cdots - l(Y_1|T_1^m)$$

$$+ l(X_1|T_1^1) + l(X_2|T_1^2) + \cdots + l(X_m|T_1^m)$$
\[
-l(Y_2|\hat{T}_1^1) - l(Y_2|\hat{T}_1^2) - \cdots - l(Y_2|\hat{T}_1^m) \\
\vdots \\
+l(X_1|\hat{T}_1^1) + l(X_2|\hat{T}_1^2) + \cdots + l(X_m|\hat{T}_1^m) \\
-l(Y_n|\hat{T}_1^1) - l(Y_n|\hat{T}_1^2) - \cdots - l(Y_n|\hat{T}_1^m)
\]
\[
= n[l(X_1|\hat{T}_1^1) + l(X_2|\hat{T}_1^2) + \cdots + l(X_m|\hat{T}_1^m)] \\
-[l(Y_1|\hat{T}_1^1) + l(Y_1|\hat{T}_1^2) + \cdots + l(Y_1|\hat{T}_1^m)] \\
+l(Y_2|\hat{T}_1^1) + l(Y_2|\hat{T}_1^2) + \cdots + l(Y_2|\hat{T}_1^m) \\
\vdots \\
+l(Y_n|\hat{T}_1^1) + l(Y_n|\hat{T}_1^2) + \cdots + l(Y_n|\hat{T}_1^m)].
\] (5.5)

This means that, roughly speaking, we need to first construct phylogenetic trees based on the jackknife sets created by all polymorphic sites. The result of this step provides \(m\) phylogenetic trees, denoted by \(\hat{T}_1^i\). Then we will evaluate \(X_i\), each polymorphic site from \(D_1\), and \(Y_j\), each polymorphic site from \(D_2\), based on \(\hat{T}_1^i\) \((i = 1, \cdots, m; j = 1, \cdots, n.)\). Totally we need to evaluate the maximum log likelihood \(m(n + 1)\) times. The computation will be extremely slow even for small values of \(m\) and \(n\). Fortunately, we may simplify this task by using approximation. Notice that inside the second \([\ ]\) of (5.5), each row is the sum of maximum log likelihood for the same unit based on \(\hat{T}_1^i\) \((i = 1, \cdots, m)\). If we view the true phylogenetic tree structure as the parameter of interest, then \(\hat{T}_1\) is the traditional estimator provided by the maximum likelihood method, and \(\hat{T}_1^i\) is the estimator identical to \(\hat{T}_1\), except that now it is computed with \(ith\) sample excluded. According to the definition of the jackknife, the average of the \(\hat{T}_1^i\)'s has the same expectation as \(\hat{T}_1\). Hence we can
replace \( l(Y_j|\hat{T}_1^i) + l(Y_j|\hat{T}_1^m) \) with \( ml(Y_j|\hat{T}_1) \) to simplify the computation. This simplification should not contradict the original intention of using \( \hat{T}_1^a \) as the estimator of \( T_1 \) because of the independence between \( \hat{T}_1 \) and \( D_2 \) by assumption. Therefore, the U-statistic we will use to estimate \( \gamma_1 \) is

\[
U_1(\hat{T}_1^i, \hat{T}_1) = \frac{1}{N} \frac{1}{m} \frac{1}{n} \left[ \sum_{i=1}^{m} l(X_i|\hat{T}_1^i) - m \sum_{j=1}^{n} l(Y_j|\hat{T}_1) \right]
\]

\[
= \frac{1}{N} \left( \sum_{i=1}^{m} l(X_i|\hat{T}_1^i)/m - \sum_{j=1}^{n} l(Y_j|\hat{T}_1)/n \right). \tag{5.6}
\]

In order to compute this statistic, only the polymorphic sites of both sets are used for calculating the maximum log likelihood. When the sites are from \( D_1 \), a technique similar to the jackknife is employed to ensure the independence between the estimated tree and the unit being evaluated. Otherwise, the tree inferred from the complete data set \( D_1 \) will be used. In other words, all or almost all the available sites are used to estimated the true phylogenetic tree. Thus the estimation can be as accurate as possible. However, only the maximum log likelihoods of the polymorphic sites are computed. This is based on the understanding that across different topologies, the maximum log likelihood for the polymorphic sites varies greatly, while it remain almost unchanged for the non-polymorphic sites. Therefore, the difference of the maximum log likelihood for the same topology between the polymorphic sites alone will provide a more clear indication of the discrepancy between the two underlining phylogenetic relationships.

To develop the final statistic for testing our hypothesis, we will apply the two-sample U-statistic theorem, which states the following:

*Let* \( X_1, \ldots, X_m \) and \( Y_1, \ldots, Y_n \) *denote independent random samples from populations with c.d.f.'s* \( F(X) \) and \( G(Y) \) *respectively. Let* \( h(\cdot) \) *be a symmetric kernel for an*
estimable parameter, \( \gamma \), of degree \((r, s)\). If \( E[h^2(X_1, \ldots, X_r; Y_1, \ldots, Y_s)] < \infty \), then

\[
\sqrt{m + n} \left[ U(X_1, \ldots, X_m; Y_1, \ldots, Y_n) - \gamma \right]
\]

has a limiting normal distribution with mean 0 and variance \( r^2 \xi_{1,0}/\lambda + s^2 \xi_{0,1}/(1 - \lambda) \) provided this variance is positive, where \( 0 < \lambda = \lim_{m+n \to \infty} (\frac{m}{m+n}) < 1 \), and \( \xi_{c,d} \) is defined as the covariance between two kernel random variables with exactly \( c \) \( X_j \)s and \( d \) \( Y_j \)s in common.

According to example 3.4.14 in Randles and Wolfe ([46]), \( \xi_{1,0} \) and \( \xi_{0,1} \) for U-statistic (5.6) here are \( \sigma^2_{1x} \) and \( \sigma^2_{1y} \), the variances of \( l(X_i|\hat{T}_1) \)s and \( l(Y_j|\hat{T}_1) \)s, respectively, and will be estimated by their corresponding sample values. It follows from the two-sample U-statistic theorem then

\[
\sqrt{m + n} \left[ U_1(\hat{T}_1, \hat{T}_1) - \gamma_1(T_1) \right] \rightarrow N(0, \frac{\sigma^2_{1x}}{\lambda} + \frac{\sigma^2_{1y}}{1 - \lambda})
\] (5.7)

provided that \( E[(l(X_i|\hat{T}_1) - l(Y_i|\hat{T}_1))^2] < \infty \) and \( \max(\sigma^2_{1x}, \sigma^2_{1y}) > 0 \).

Similarly, the U-statistic to estimate \( \gamma_2 = \frac{l(D_2|T_2) - l(D_1|T_2)}{N} \) is:

\[
U_2(\hat{T}_2, \hat{T}_2) = \frac{1}{N} \left( \sum_{j=1}^{n} l(Y_j|\hat{T}_2)/n - \sum_{i=1}^{m} l(X_i|\hat{T}_2)/m \right)
\] (5.8)

and its asymptotic distribution is also given by

\[
\sqrt{m + n} \left[ U_2(\hat{T}_2, \hat{T}_2) - \gamma_2(T_2) \right] \rightarrow N(0, \frac{\sigma^2_{2x}}{\lambda} + \frac{\sigma^2_{2y}}{1 - \lambda})
\] (5.9)

where \( \sigma^2_{2x} \) and \( \sigma^2_{2y} \) are the variances of \( l(X_i|\hat{T}_2) \)s and \( l(Y_j|\hat{T}_2) \)s, respectively. They will be estimated by their sample values as well. This leads to the main result of this chapter.

**Theorem 5.1.1** Let \( X_1, \ldots, X_m \) and \( Y_1, \ldots, Y_n \) be sample polymorphic sites among the \( u \) and \( v \) sites from \( D_1 \) and \( D_2 \) respectively, \( T_1 \) and \( T_2 \) be the topologies of the
true (unknown) phylogenetic trees of $D_1$ and $D_2$, respectively. $B$ is used to stand for branch lengths in the tree. Throughout the statement and proof of this theorem, we will use $\ln(X|T, B)$ to denote the log likelihood of data $X$ for given tree topology $T$ and branch lengths $B$.

Furthermore, we assume the maximum log likelihood values

$$\max_B \frac{\ln(X_1|T_i, B)}{N}, \ldots, \max_B \frac{\ln(X_m|T_i, B)}{N}$$

are i.i.d. with mean $m_X(T_i)$ and variance $\sigma_X^2(T_i) < \infty$;

$$\max_B \frac{\ln(Y_1|T_i, B)}{N}, \ldots, \max_B \frac{\ln(Y_n|T_i, B)}{N}$$

are i.i.d. with mean $m_Y(T_i)$ and variance $\sigma_Y^2(T_i) < \infty$, where $(i = 1, 2)$.

Also assume that as $m \to \infty$, and $n \to \infty$

$$\frac{m}{n} \to c > 0$$

Also let

$$\gamma_1(T_1) = m_X(T_1) - m_Y(T_1)$$

$$\gamma_2(T_2) = m_Y(T_2) - m_X(T_2)$$

Then

$$\frac{U_1(\hat{T}_1, \hat{T}_1) - \gamma_1(T_1)}{\sqrt{\frac{\sigma_X^2(T_1)}{m} + \frac{\sigma_Y^2(T_1)}{n}}} \quad (5.10)$$

and

$$\frac{U_2(\hat{T}_2, \hat{T}_2) - \gamma_2(T_2)}{\sqrt{\frac{\sigma_X^2(T_2)}{n} + \frac{\sigma_Y^2(T_2)}{m}}} \quad (5.11)$$

both have an asymptotically standard normal distribution, where $U_1(\hat{T}_1, \hat{T}_1)$ and $U_2(\hat{T}_2, \hat{T}_2)$ are defined in (5.6) and (5.8), respectively.
Proof:

Because of the symmetric structure, it is sufficient to prove (5.10) is asymptotically standard normal distributed only. Throughout the proof and the rest of this chapter, we will use \( l(X|T) \) as the convenient notation for \( \max_B(\ln(X|T, B)) \).

Here we only assume that both \( m \) and \( n \) approach infinity. But since the polymorphic sites are only portions of the entire sequences, it indicates that both \( u \) and \( v \), the lengths of \( D_1 \) and \( D_2 \) respectively, should approach infinity as well.

As described before, the U-statistic used to estimate \( \gamma_1(T_1) \) is

\[
U_1(T_1, T_1) = \frac{1}{N} \left( \sum_{i=1}^{m} l(X_i|T_1)/m - \sum_{j=1}^{n} l(Y_j|T_1)/n \right). \tag{5.12}
\]

Under the mild conditions described before, by the central limit theorem for two sample U-statistics, as \( m \to \infty \), and \( n \to \infty \)

\[
\frac{U_1(T_1, T_1) - \gamma_1(T_1)}{\sqrt{\frac{\sigma_X^2(T_1)}{m} + \frac{\sigma_Y^2(T_1)}{n}}} \Rightarrow N(0, 1). \tag{5.13}
\]

We use \( \Rightarrow \) to indicate convergence in distribution. It is a well-known fact that (5.13) remains true if the unknown population variances \( \sigma_X^2(T_1) \) and \( \sigma_Y^2(T_1) \) are replaced by their corresponding sample variances \( \hat{\sigma}_X^2(T_1) \) and \( \hat{\sigma}_Y^2(T_1) \), respectively.

However, there still exists one difficulty for the approach to be useful. Namely, \( T_1 \) is unknown in general. It is shown by Chang ([4]) and others that the maximum likelihood method is consistent for constructing a phylogenetic tree. Therefore if we let \( \hat{T}_1 \) denote the maximum likelihood estimator of \( T_1 \) based on the the entire data set of \( X_1, \ldots, X_k \), \( \hat{T}_1^i \) be the same estimator but with \( i \)th polymorphic site removed, we know that for \( i = 1, \ldots, m \), when \( u \to \infty \)

\[
\hat{T}_1 \xrightarrow{p} T_1 \\
\hat{T}_1^i \xrightarrow{p} T_1.
\]
Since $\ln(X|T_1,B)$ is a continuous function of $B$ for given $X$ and $T_1$, so is its maximized function. Therefore,

$$l(X_i|\hat{T}_{i1}) \xrightarrow{P} l(X_i|T_1). \quad (5.14)$$

Similarly

$$l(Y_j|\hat{T}_1) \xrightarrow{P} l(Y_j|T_1) \quad (5.15)$$

where $i = 1, \ldots, m, j = 1, \ldots, n$.

Furthermore, $X_i$s and $Y_j$s are i.i.d. random variables by assumption. Therefore, the convergence in (5.14) and (5.15) is uniform: $\forall \epsilon_1 > 0, \epsilon_2 > 0, \exists N_1 = N(\epsilon_1), N_2 = N(\epsilon_2)$, such that $\forall i, j$

when $u \geq N_1$

$$|l(X_i|\hat{T}_{i1}) - l(X_i|T_1)| < \epsilon_1,$$

when $v \geq N_2$

$$|l(Y_j|\hat{T}_1) - l(Y_j|T_1)| < \epsilon_2.$$

Since the values of $N_1$ and $N_2$ only depend on $\epsilon_1$ and $\epsilon_2$, respectively, it is straightforward to show that

when $u \geq N_1$

$$\left| \frac{\sum_{i=1}^{m} l(X_i|\hat{T}_{i1})}{m} - \frac{\sum_{i=1}^{m} l(X_i|T_1)}{m} \right| < \frac{m\epsilon_1}{m} = \epsilon_1 \quad (5.16)$$

when $v \geq N_2$

$$\left| \frac{\sum_{j=1}^{n} l(Y_j|\hat{T}_1)}{n} - \frac{\sum_{j=1}^{n} l(Y_j|T_1)}{n} \right| < \frac{n\epsilon_2}{n} = \epsilon_2. \quad (5.17)$$
Equations (5.16) and (5.17) indicate that

\[
\frac{\sum_{i=1}^{m} l(X_i|\hat{T}_i)}{m} - \frac{\sum_{i=1}^{m} l(X_i|T_1)}{m} \xrightarrow{P} 0 \quad (5.18)
\]

\[
\frac{\sum_{j=1}^{n} l(Y_j|\hat{T}_1)}{n} - \frac{\sum_{j=1}^{n} l(Y_j|T_1)}{n} \xrightarrow{P} 0. \quad (5.19)
\]

Therefore,

\[
U_1(\hat{T}_1, \hat{T}_1) - U_1(T_1, T_1)
\]

\[
= \frac{1}{N} \left( \frac{\sum_{i=1}^{m} l(X_i|\hat{T}_i)}{m} - \frac{\sum_{i=1}^{m} l(X_i|T_1)}{m} \right) - \frac{1}{N} \left( \frac{\sum_{j=1}^{n} l(Y_j|\hat{T}_1)}{n} - \frac{\sum_{j=1}^{n} l(Y_j|T_1)}{n} \right)
\]

\[
\xrightarrow{P} 0. \quad (5.20)
\]

Similarly, one can also show that, as \( u \to \infty, v \to \infty \)

\[
\frac{\partial^2_{\chi}(\hat{T}_i)}{\partial^2_{\chi}(T_1)} - \frac{\partial^2_{\chi}(\hat{T}_1)}{\partial^2_{\chi}(T_1)}
\]

\[
= \frac{1}{N^2} \left( \frac{\sum_{i=1}^{m} (l(X_i|\hat{T}_i)/m)^2}{m} - \frac{\sum_{i=1}^{m} (l(X_i|T_1)/m)^2}{m} \right) - \frac{1}{N^2} \left( \frac{\sum_{j=1}^{n} (l(Y_j|\hat{T}_1)/n)^2}{n} - \frac{\sum_{j=1}^{n} (l(Y_j|T_1)/n)^2}{n} \right)
\]

\[
\xrightarrow{P} 0
\]

which indicate that

\[
\frac{\partial^2_{\chi}(\hat{T}_i)/m}{\partial^2_{\chi}(T_1)/m} = \frac{\partial^2_{\chi}(\hat{T}_1)}{\partial^2_{\chi}(T_1)} \xrightarrow{P} 1 \quad (5.21)
\]

and

\[
\frac{\partial^2_{\chi}(\hat{T}_1)/n}{\partial^2_{\chi}(T_1)/n} = \frac{\partial^2_{\chi}(\hat{T}_1)}{\partial^2_{\chi}(T_1)} \xrightarrow{P} 1. \quad (5.22)
\]
From (5.13) and some simple algebra, we know that

\[ \frac{U_1(T_1, T_1) - \gamma_1(T_1) + (U_1(\hat{T}_1, \hat{T}_1) - (U_1(\hat{T}_1, \hat{T}_1)))}{\sqrt{\frac{s^2(T_1)}{n} + \frac{s^2(\hat{T}_1)}{m}}} \times \sqrt{\frac{s^2(T_1)/m + s^2(\hat{T}_1)/n}{s^2(\hat{T}_1)/m + s^2(\hat{T}_1)/n}} \]

Similarly,

\[ \frac{U_1(\hat{T}_1, \hat{T}_1) - \gamma_1(T_1) + (U_1(T_1, T_1) - (U_1(\hat{T}_1, \hat{T}_1)))}{\sqrt{\frac{s^2(T_1)}{m} + \frac{s^2(\hat{T}_1)}{n}}} \times \sqrt{\frac{s^2(T_1)/m + s^2(\hat{T}_1)/n}{s^2(T_1)/m + s^2(\hat{T}_1)/n}} \]

\[ \Rightarrow \ N(0, 1). \quad (5.23) \]

Combining this with equations (5.20), (5.21), and (5.22), we complete the proof by applying Slutsky's theorem to obtain the desired result

\[ \frac{U_1(\hat{T}_1, \hat{T}_1) - \gamma_1(T_1)}{\sqrt{\frac{s^2(\hat{T}_1)}{m} + \frac{s^2(\hat{T}_1)}{n}}} \Rightarrow \ N(0, 1). \quad (5.24) \]

Similarly,

\[ \frac{U_2(\hat{T}_2, \hat{T}_2) - \gamma_2(T_2)}{\sqrt{\frac{s^2(\hat{T}_2)}{n} + \frac{s^2(\hat{T}_2)}{m}}} \Rightarrow \ N(0, 1). \quad (5.25) \]

Notice that under the null hypothesis, \( \gamma_1(T_1) = \gamma_2(T_2) = 0 \). Therefore, when both \( D_1 \) and \( D_2 \) have enough polymorphic sites, two statistics

\[ Z_1 = \frac{U_1(\hat{T}_1, \hat{T}_1)}{\sqrt{\frac{s^2(\hat{T}_1)}{m} + \frac{s^2(\hat{T}_1)}{n}}} \]

\[ Z_2 = \frac{U_2(\hat{T}_2, \hat{T}_2)}{\sqrt{\frac{s^2(\hat{T}_2)}{n} + \frac{s^2(\hat{T}_2)}{m}}} \]

(5.26)

(5.27)

can be used for the significance test, with \( U_1 \) and \( U_2 \) being the two statistics used to estimate the parameters \( \gamma_1 \) and \( \gamma_2 \), respectively. It appears that, however, they can not be combined together to form a single statistic for testing the hypothesis of interest. The maximum likelihood method has been found to be consistent for
finding the phylogenetic tree topology. In other words, the maximum likelihood method will eventually find the true tree topology when the sample size (length of sequences) becomes large enough. Therefore, under the null hypothesis that the two sets of sequences are related by the same phylogenetic tree, \( \hat{T}_1, \hat{T}_2, \hat{T}_i, \) and \( \hat{T}_i' \) will all provide accurate and approximate identical estimation of the true phylogenetic tree, in particular for the consideration of topology. Therefore statistics \( U_1 \) (5.6) and \( U_2 \) (5.8) are highly negatively correlated. Approximately, for large enough samples, \( U_1 = -U_2 \). In other words, intuitively, \( U_1 \) should provide about the same information as \( U_2 \) does when the null hypothesis is true. Hence, it is sufficient to use only one of these two statistics if one is interested in detecting if the phylogenetic trees connecting the two sets of species are the same. In the view of a technical point, it may be difficult to find a statistic which utilizes both \( U_1 \) and \( U_2 \) even if one wishes to do so because of the high correlation between them. One possible statistic to combine both \( U_1 \) and \( U_2 \) is

\[
Z' \Sigma^{-1} Z \Rightarrow \chi_k^2 
\]  

(5.28)

where

\[
Z = \begin{bmatrix} Z_1 \\ Z_2 \end{bmatrix} = \begin{bmatrix} \frac{U_1(\hat{T}_1, \hat{T}_1)}{\sqrt{s^2_{m}(\hat{T}_1) + s^2_{n}(\hat{T}_1)/n}} \\ \frac{U_2(\hat{T}_2, \hat{T}_2)}{\sqrt{s^2_{m}(\hat{T}_2) + s^2_{n}(\hat{T}_2)/m}} \end{bmatrix}
\]

\[
\Sigma = \text{cov}(Z) = \begin{bmatrix} 1 & \text{cor}(Z_1, Z_2) \\ \text{cor}(Z_1, Z_2) & 1 \end{bmatrix}
\]

and

\[
k = \text{rank}(\Sigma).
\]
Under the null hypothesis, both $Z_1$ and $Z_2$ are asymptotically standard normal distributed. The high correlation between $U_1$ and $U_2$ leads to high correlation between $Z_1$ and $Z_2$. The statistic $Z'\Sigma^{-1}Z$ of (5.28) becomes very sensitive because the estimation of $\Sigma^{-1}$ involves the estimation of $\frac{1}{1-\text{cor}(Z_1,Z_2)^2}$, whose denominator can be extremely close to 0. This will make the statistic proposed in (5.28) an unstable, and therefore unreliable test statistic. When the alternative is true, however, $U_1$ and $U_2$ are less strongly correlated as expected, though in general their correlation is still likely to be negative. Hence we recommend to use only one of these two statistics when testing the hypothesis. However, if one wishes to construct confidence intervals for the two parameters, $\gamma_1$ and $\gamma_2$, both $U_1$ and $U_2$ need to be computed.

5.2 Computer Implementation and Approximation

To implement our method on the computer, we employ programs from PHYLIP, a package of computer programs for inferring phylogenies written by Joseph Felsenstein, to construct the inferred trees based on the nucleotide sequences and evaluate the maximum log likelihood of each single site of these sequences based on an estimated tree. As discussed in the previous section, the computation may be very extensive for two reasons: first, the maximum likelihood methods in phylogenetic study are generally very slow, even for just a few species; secondly, to eliminate the dependence between the estimated tree and the site being evaluated, we use a technique that is similar to the Jackknife: the estimated tree is constructed from the sequences with each single site removed. For example, for a set of $N$ sequences with $n$ polymorphic sites, the proposed method requires us to construct phylogenetic trees $n$ times, each time with a single polymorphic site removed from the original nucleotide sequences.
Hence, it is usually very difficult to do the computation when \( n \) and \( N \) are large.

It is clear that an approximation should be used to improve the efficiency of the method. In the original approach, the maximum likelihood method was used to both construct the inferred tree and to evaluate the likelihood of a single site. While it may be essential to use the maximum likelihood method for the second task, other methods, such as the distance matrix method, which is much faster, may be suitable for constructing the estimated trees.

Distance matrix methods fit a tree to a matrix of pairwise distance between the species. For nucleotide sequence data the distance may, for example, be calculated from the fraction of sites different between the two sequences. And a phylogenetic tree is constructed by using an algorithm based on some functional relationships among the distance values. Fitch and Margoliash ([13]) introduced the first distance matrix method. Other widely used methods include Farris’s “distance Wagner method” ([9]), Li’s method ([35]), Tateno et al.’s “modified Farris method” ([60]), and Saitou and Nei’s “neighbor joining method” ([50]).

5.3 Simulation Study

In the previous sections, we have developed a computational feasible method to test if two sets of nucleotide sequences are sampled from an identical phylogenetic tree. If the relations among these two sets indeed are the same, the statistic we proposed should have an asymptotically standard normal distribution. Otherwise, the distribution of the statistic shall roughly remain bell shaped but be shifted to right by a certain amount which depends on the difference between the two true phylogenies from which the nucleotide sequences are generated. Though the properties
of the statistic have been studied analytically, it remains very important to see how accurate they are for different situations.

In this section, computer simulation studies will be conducted to assess the performance of the statistics we proposed. As can be seen from theorem (5.1.1), the asymptotic normality of the test statistic relies on having a large number of polymorphic sites. This simulation study shall focus on investigating this relationship. Another issue, the efficacy of different tree-making methods, however, will not be included since extensive research has been done by others (see Felsenstein[11] for review). Thus, the sample sizes of this study are not the total numbers of nucleotide sites. Instead, they are the numbers of the polymorphic sites in the two sets. Because of the natural difference between the null hypothesis and the alternative hypothesis, separate evaluation will be done accordingly.

5.3.1 Approximation

One concern for this approach is the slowness of the maximum likelihood method. As discussed in section 5.2, an approximation is proposed to resolve the difficulty: we suggest using a computationally faster method for constructing the inferred tree. Here we shall use the distance matrix method to accomplish this task.

One natural question asks how good is this approximation. Our simulation studies will first try to address this question when the null hypothesis is true. The required computing time for the maximum likelihood method increases rapidly with the number of species ($N$). It becomes practically impossible to conduct the computer simulation when $N$ is large. We are only able to simulate the situation of
$N = 6, m = n = 100$ (i.e., each set has 100 polymorphic sites) to assess the approximation.

In the simulation, under the null hypothesis, 100 pairs of nucleotide data sets $(D_1, D_2)$ are generated from the same phylogenetic trees. The statistic (5.26) is evaluated for each pair by two slightly different approaches: 1. the original one using maximum likelihood method ($ML$); 2. the one using distance matrix method ($DS$). The results are then compared to find out how accurate the approximation is.

Figure (5.1) shows the approximate distribution of statistic (5.26) for the two methods we discussed based on 100 simulations. The general patterns in both graphics are very similar: the obtained histograms are reasonably close to a standard normal distribution.
curve. It is therefore suggested that statistic (5.26) is approximately a standard normal distribution regardless which approach is used. However, in this study, instead of the asymptotic behavior of the statistic itself, we are more interested in measuring the difference between the two approaches. This question can be answered more clearly by Figure (5.2). The histogram illustrates the effect of the approximation on statistic (5.26). The difference between $Z(ML)$ and $Z(DS)$ distributes around 0, with very small deviation. It indicates that although the approximation may change the value of (5.26) from the original approach, the difference, however, is very small with the majority less than 0.1, and the largest being about 0.2. Besides, the differences do not seem to be systematic since both histograms evenly spread around the center. Overall, it is convincing that the approximation is acceptable according to the results.
of our simulation.

Computer simulations have also been conducted by others (e.g. Saitou and Imanishi [49], Tateno etc. [61]) to evaluate the efficiencies of several tree-making methods for obtaining the correct phylogenetic tree. Though the maximum likelihood method is slightly better when the evolutionary rate varied drastically among branches, it is found that in general the maximum likelihood method and a distance matrix method which uses corrected nucleotide substitutions show more or less the same performance. Topology estimates are shown to be more robust than branch length estimates of a tree with respect to violations of the assumption of the mathematical model used for both methods.

Because of the reasonable accuracy of the approximation, especially when the tree topology is the parameter of interest, together with the fact that distance matrix methods are much faster than the maximum likelihood method, we will employ them in the evaluation of (5.26). Throughout this chapter, the $Z(\mathcal{DS})$ version of (5.26) is used unless otherwise indicated.

5.3.2 Results when the null hypothesis is true

The most important property of a test statistic is its ability to reject the alternative hypothesis when the null hypothesis is true at the appropriate error rate. In our case, it means that when the two data sets under study are generated from the same phylogenetic tree, then the statistic should be able to draw such a conclusion with a specified high probability. To be more precisely, the statistic we proposed should approximately follow a standard normal distribution. Another conclusion we
will also investigate is that the two statistics (5.26) and (5.27) are asymptotically identical when the null hypothesis is true. The difference between these two statistics is that they are evaluated based on different inferred trees. The tree used in $Z_1$ is estimated from $D_1$, while the tree used in $Z_2$ is from $D_2$. Under the null hypothesis that $T_1 = T_2$, as both $m$ and $n$ increase, (and so do $u$ and $v$) $\hat{T}_1 \approx \hat{T}_2$, and $\hat{T}_2 \approx \hat{T}_1$ for all $i$ and $j$. Consequently, $Z_1 \approx -Z_2$. Therefore, when testing the hypothesis, we only need to compute one of these two statistics. We will conduct simulation studies to assess these two conclusions.

One question that may also be asked is whether having an unequal number of polymorphic sites (i.e., $m \neq n$) will affect these two conclusions. During the theoretical deduction in the previous section, the only assumption regarding the two sample sizes is that as both $m$ and $n$ become larger and larger, their ratio approaches a nonzero constant. Therefore, the asymptotic behavior of $Z_1$ and $Z_2$ for unequal sample sizes should be the same as those for equal sample sizes. However, it remains an interesting issue in practice where the sample sizes are finite. This question will also be answered by the computer simulation studies.

Two sets of $N$ nucleotide sequences are first generated according to the same topology, one has $m$ polymorphic sites and the other has $n$. To study the effect of

<table>
<thead>
<tr>
<th>N</th>
<th>m=100</th>
<th>n=100</th>
<th>n=500</th>
</tr>
</thead>
<tbody>
<tr>
<td>N=6</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>N=16</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>N=32</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Table 5.2: Values of $N$, $m$, and $n$ in the simulation study
the number of species \((N)\) and the number of polymorphic sites \((m, n)\) on the performance of the statistics, several combinations of \((N, m, n)\) will be used to produce the data. We set \(N = 6, 16, \text{ and } 32\), and for each \(N\), \((m, n)\) takes the values of \((100, 100)\), \((100, 500)\), and \((500, 500)\) as indicated in table (5.2).

![Tree diagram]

Figure 5.3: Topology used to generate sequences for \(N = 6\) under null hypothesis

Because the sample size here is the number of polymorphic sites, the number of species does not seem to have a direct relation with the asymptotic behavior of the statistic. However, since it is not completely clear how the number of species affects the accuracy of the inferred tree, which in turn may affect the test statistic, it remains interesting to see how robust the statistic is regarding this matter. The trees used to generated the sequences are shown in figures (5.3), (5.5), and (5.7).
Figure 5.4: Results of 100 simulations of statistics $Z_1(5.26)$ and $Z_2(5.27)$ for $N=6$, the histograms on the first row are based on $m = n = 100$; the second row are based on $m = 100$ and $n = 500$; and the third row are based on $m = n = 500$.

The results for $N = 6$ are presented in figure (5.4). The histograms on each row are associated with a particular combination of $m$ and $n$. The rightmost histogram on each row is based on the sum of the two statistics. We draw the following conclusions based on these figures:

1. Both statistics $Z_1$ and $Z_2$ are reasonably close to a standard normal distribution when the minimum of $m$ and $n$ is at least 100 sites;
2. The simulation results agree with the earlier analytic conclusion that $Z_1$ is approximately the opposite of $Z_2$. The most obvious evidence is when both $m$ and $n$ are 500, $Z_1$ becomes exactly the opposite of $Z_2$;

3. From the graphics based on the results of $m = 100$, and $n = 500$, we conclude that there is not much difference between the distribution of $Z_1$, which is evaluated on the tree inferred from the data set with fewer polymorphic sites, and $Z_2$, which uses the data set with more polymorphic sites.
Figure 5.6: Results of 100 simulations of statistics $Z_1(5.26)$ and $Z_2(5.27)$ for $N=16$, the histograms on the first row are based on $m = n = 100$; the second row are based on $m = 100$ and $n = 500$; and the third row are based on $m = n = 500$.

The results for $N = 16$, and $N = 32$ are also presented in the figures (5.6) and (5.8). A similar pattern is found across different values of $N$. Though for $N = 16$ and $N = 32$, $Z_1$ and $Z_2$ are not exactly the opposite of each other even when $m = n = 500$ as for the case of $N = 6$. However, the difference is very negligible. It is therefore suggested that under the null hypothesis the simulation results coincide with the conclusions of the analytic analysis. Some of those conclusions are that statistics $Z_1$
Figure 5.7: Topology used to generate sequences for $N = 32$ under null hypothesis

and $Z_2$ provide about the same amount of information and that evaluation of the statistics' distribution do not depend on the number of polymorphic sites as long as the number is reasonably large. Based on this finding, only the statistic $Z_4$ for the case of equal numbers of polymorphic sites is included in the simulation study under the alternative hypothesis.
Figure 5.8: Results of 100 simulations of statistics $Z_1(5.26)$ and $Z_2(5.27)$ for $N=32$, the histograms on the first row are based on $m = n = 100$; the second row are based on $m = 100$ and $n = 500$; and the third row are based on $m = n = 500$.

### 5.3.3 Results when the alternative hypothesis is true

Another goal of the simulation studies is to assess the power of the statistic: the probability of rejecting the null when the alternative hypothesis is true. It is more complicated than the case under the null hypothesis because numerous situations can lead to the violation of the null hypothesis. Obviously it is impossible to reproduce each of these possibilities. Our plan will be to focus on investigating the relationships
between the value of the statistic and the difference of the two phylogenetic trees: As
the difference changes, how does the test statistic adjust with it? When there exists
no difference, it reduces to the simpler case of null hypothesis.

Our immediate objective is to find a series of phylogenetic trees, such that each tree
is different from a common tree by a distinct amount. A set of nucleotide sequences
generated from the common tree is then paired with each set of sequences produced
from this series of phylogenetic trees. The statistic developed in the previous section
will be applied to every pair of these sets of sequences for detecting the discrepancy
between the pair of the phylogenies. As discussed earlier, the asymptotic distribution
of the statistic under the alternative hypothesis is only different from that under the
null hypothesis by a shift of location which is associated with the difference between
the two phylogenies under study. The degree of discordance is then compared among
all the pairwise comparisons. Computer simulations are conducted 100 times for each
pair of phylogenies for these comparisons. The parameter $\gamma_1$ (5.1) is 0 under the null
hypothesis and is larger than 0 otherwise. The hypothesis test is therefore a one­

sided test. Consequentially, the null hypothesis will be rejected at the $\alpha$ level of 0.05
if the statistic yields a value bigger than 1.645. The percentages of simulations for
which a significant difference is found between these pairs of phylogenies are recorded.
These percentages are considered the approximate power of the statistic. As with the
simulation study under the null hypothesis, phylogenetic trees consisting of 6, 16,
and 32 species are included in this study. And the sequences of different lengths will
again be used to study the asymptotic behavior of the statistic.
For the case of $N = 6$, the phylogenies used to generate the nucleotide sequences are listed in figure (5.9). The leftmost tree on the first row is regarded as the common tree in our study. The trees that are closer to the common tree are listed on the first row. Across the same row, trees are arranged from left to right according to their difference from the common tree. The tree at the right of the second row is the one which has the largest difference from the common tree. A set of 6 nucleotide sequences is generated from each of these 6 phylogenetic trees. Again, sequences of 100 and 500 polymorphic sites are used to evaluate the effect of the sample size on the power of the statistic.
Estimated power curves of the statistic (5.26) are shown in figure (5.10). It illustrates the relationships between the values of $\gamma_1$ (5.1) and the power of the statistic. The estimates under the null hypothesis, i.e., when the difference between the two phylogenies is 0, are also included to provide a more complete analysis on the behavior of the statistic. Results from sequences of different lengths are presented in the same graph to see the effect of the sample size.

![Figure 5.10: Estimated power curves for $N = 6$](image)

As can be seen from Figure (5.10), the estimated power increases with the parameter $\gamma_1$ as expected. The test is more likely to correctly reject the null hypothesis as the distance between the null and alternative values increases. The effect of the sample size is very obvious since the estimated power curve for $m = n = 500$ is clearly
above that for \( m = n = 100 \).

Parallel simulations are also done for the case of \( N = 16 \) and \( 32 \). The phylogenetic trees (5.5) and (5.7) are used as the common trees for the case of 16 and 32 sequences, respectively. Two series of phylogenies are formed by making changes from these common trees. The phylogenetic trees of both series will be increasingly different from the corresponding common tree. We start with 16 species from Figure (5.5).

1. Species2 switches with species3 to form the second phylogeny;

2. Species2 switches with species7 to form the third phylogeny;

3. Species2 switches with species3, and species6 switches with species8 to form the fourth phylogeny;

4. Species8 switches with species9 to form the fifth phylogeny;

5. Species2 switches with species3, and species8 switches with species9 to form the sixth phylogeny.

And for the case of 32 species, we start from Figure (5.7).

1. Species2 switches with species3 to form the second phylogeny;

2. In the second phylogeny, species5 switches with species8 to form the third phylogeny;

3. In the third phylogeny, species9 switches with species11 to form the fourth phylogeny;
4. In the third phylogeny again, species 18 switches with species 19 to form the fifth phylogeny;

5. In the fourth phylogeny, species 18 switches with species 19, species 22 switches with species 23, and species 26 switches with species 27 to form the sixth phylogeny.

Figure 5.11: Estimated power curves for $N = 16$ and 32
After the two series are completed, nucleotide sequences with 100 and 500 polymorphic sites are generated from each of these phylogenies. The results of the studies are shown in Figure (5.11). Similar patterns are found from these graphs: the power increases with both sample size and distance between the null and alternative parameters. One question that has been mentioned at the beginning of the simulation study is how the test statistic (5.26) will react to the increasing of the number of sequences. Though all three curves appear to be similar, the power increases more rapidly when the number of sequences under study is larger. This is indicated by the fact the test statistic (5.26) with more sequences achieves the same power within a shorter distance when we compare the ranges of the X-axes of all three plots. It is therefore suggested that a certain number of sequences with sufficient polymorphic sites be essential for obtaining adequate power in applying this method.

5.4 Application

In this chapter, a statistical method has been developed to detect the discrepancy between two phylogenies. Computer simulations have shown that this method is highly effective. The method is useful for many purposes. Here we will apply it to reexamine the existence of a recombination event.

Acquired Immune Deficiency Syndrome (AIDS) is caused by the human immunodeficiency virus which has major subtypes HIV-1 and HIV-2. HIV is a member of the lentivirus family of retroviruses. A growing number of related viruses have also been discovered in nonhuman primates: these have been termed simian immunodeficiency viruses (SIV). Both HIV and SIV encode several genes, including gag, pol, and env. HIVs and SIVs exhibit extraordinary genetic diversity. This is due to their
very high rate of nucleotide sequence evolution and recombination. Recombination events can be identified most clearly in the context of phylogenetic analysis. The existence of a recombination in the genome is indicated by the discordance of phylogenetic relationships for different genes or different regions of the same gene. These discrepancies can be found by applying the method we have proposed.

First it is necessary to review briefly the current state of knowledge concerning the phylogenetic relationships among HIV-1. There are two distinct groups of HIV-1 viruses. All of the earliest-known HIV-1 isolates fall into one of these clusters, called group M. Phylogenetic analyses of viruses collected from all over the world have revealed at least eight subgroups within this cluster, termed sequence subtypes A-H ([55]). The second major cluster, called group O, has been identified only recently ([17]) and is known only from a small number of isolates originating from Cameroon. Recombination among HIV viruses can be detected when different genes, or different regions within the same gene, are placed by phylogenetic analysis into different sequence subtypes.

The MAL strain of HIV-1 was one of the first of African origin to be characterized. MAL was isolated from a boy with AIDS-related complex in Zarie. It has long been suspected to be recombinant. Li et al. ([38]) first performed a phylogenetic analysis of env gene sequences from the relatively limited number of American and African isolates of HIV-1 available at that time. MAL was found to be more closely related to other African isolates (ELI, member of what is now known as sequences subtype D) than to the American isolates (member of subtype B). Since then, the sequences of a much larger number of HIV-1 isolates have been determined. Phylogenetic analyses of env sequences have MAL falling in subtype D. Parallel analyses of gag sequences
have MAL clustering with subtype A virus ([55]). Here we shall apply our method to reexamine the mosaic structure of the sequences.

The aligned sequences of both \textit{gag} and \textit{env} genes were downloaded from the HIV sequences database: The Human Retroviruses and AIDS 1995 Compendium. In total twelve sequences are included in this study. They are U455 and DJ258 of subtype A, Lai, MN, and OYI of subtype B, SM145 of subtype C, ELI, Z2Z6, and NDK of subtype D, BZ163 of subtype F, ANT70 of group O served as an outgroup, and MAL as the recombinant.

Figure 5.12: The mosaic genome of MAL of HIV-1. The left tree is constructed from the \textit{gag} gene, and the right tree is from the \textit{env} gene.
Figure (5.12) depicts the estimated phylogenetic trees of *gag* and *env* genes by using programs in the PHYLIP package. It shows that the *gag* gene of MAL is most closely related to that of a representative member of subtype A (U455), and the *env* gene of MAL is most closely related to subtype D viruses NDK, Z2Z6, and ELI. It thus suggests that different regions of the genome of MAL have different phylogenetic histories. However, does it indicate that MAL, or an ancestor of MAL, has resulted from recombination? Is this appeared mosaic structure of MAL purely due to chance?

We have to remember that the conclusion of recombination should be, ideally, drawn based on the difference of the true phylogenetic relationships, but not that of the estimated ones. This problem can be solved by applying the method we obtained in this chapter.

With 182 polymorphic sites from *gag* genes and 193 polymorphic sites from *env* genes, the proposed statistic (5.26) has a value of 1.854486, which in turn yields a p-value of 0.032 for a one-sided test. Therefore we draw the conclusion that the two underlying phylogenetic relationships of *gag* and *env* genes indeed are statistically different. Consequentially, this suggests that there has been a recombination event in the regions between *gag* and *env*. 
Conclusions

The study of molecular evolution has its roots in two disparate disciplines: population genetics and molecular biology. Population genetics provides the theoretical foundation for the study of evolutionary processes, while molecular biology provides the empirical data. The chromosomal or genomic location of a gene is called a locus, and alternative forms of the gene at a given locus are called alleles. The substitution of allele in a population generally takes thousands or even millions of years to complete. For this reason, we cannot deal with the process of molecular substitution by direct observation, and nucleotide substitutions are always inferred from pairwise comparisons of DNA molecules that share a common origin. Changes in molecular sequences are used in molecular evolutionary studies both for estimating the rate of evolution and for reconstructing the evolutionary history of organisms.

Nucleotide mutation has long been considered as a very important factor in evolution. On one hand, a redundant duplicate of nucleotides may acquire divergent mutations and eventually a new gene emerges. However, this change is generally very slow because this process is restricted by the biological function of a protein, the product of the genes. This is especially true if the change is not functionally desirable. Other means of creating new functions are now known. It is suggested that
recombination provides a faster mechanism for evolution to create new functions.

It is known that the rate of synonymous substitution, which does not change the amino acid, is generally much greater than that of nonsynonymous substitution, which causes an amino acid change. Though the rate of synonymous substitution varies considerably among genes, it is much less than the nonsynonymous rate. Thus synonymous substitutions may be used as a molecular clock for dating the evolutionary time of closely related species. Furthermore, the comparison between synonymous and nonsynonymous substitution rates may be used as an indicator of the intensity of the selective constraints on the amino acid sequence of a protein. Therefore, in the study of the evolutionary divergence of DNA sequences, it is often required to estimate the numbers of synonymous and nonsynonymous nucleotide substitutions separately, along with their ratio.

Though many estimators (see chapter 2) for estimating the synonymous and nonsynonymous substitution rate, together with their corresponding variances, have been proposed, no currently used method assesses the variation of the ratio of the synonymous to nonsynonymous substitution rates. In chapter 3, we developed a new method to handle this problem. It is based on the pattern of codon degeneracy in the DNA code (Table 1.1). Two slightly different versions of this method are presented, one adopts Jukes and Cantor’s ([29] one-parameter model and the other adopts Kimura’s ([32]) two-parameter model. Computer simulation indicates that the estimates by the new method are very accurate and comparable to those given by Li’s ([36]) improved method when the comparison is available.

A phylogenetic tree describes the pattern of historical relationships among a group of individuals, genes, or species. It can be reconstructed using a variety of methods,
usually maximum likelihood, distance matrix, or maximum parsimony. As long as the
genomes of interest have remained intact and unmixed, the same pattern of branching
will recount the ancestry of every region of the genome. This is the basic idea behind
the several applications of phylogenetic analyses. One of these applications is the
discovery of recombination between divergent viruses. With recombination within a
species, different genes or different regions of the same gene would generally lead to
different phylogenetic trees. However, the inferences made from an estimated tree
about the evolution of a group depend on the accuracy of the estimated tree (see
chapter 4 for more review). Therefore, it is necessary to account for the variability
in the two inferred trees when one tries to make a statement about any difference
between them.

In chapter 5, we proposed a statistic for detecting the difference between two
underlying phylogenetic trees from the observed nucleotide sequences. The statistic
utilizes the difference of maximum log likelihood for the same topology between
the two sources as an indicator of the discordance between the two true relationships.
Thus the effect of the tree estimation is reduced. To ensure the independence between
the inferred tree and the nucleotide sites being evaluated, a leave-one-out procedure
is employed. When the two true trees have the same topology, the statistic follows,
asymptotically, a standard normal distribution. Therefore, this issue can be tested
statistically.

In summary, two very important issues of the evolutionary process are approached
in a manner that takes into account the uncertainty of the estimates. A general
method is developed to test whether a group of genes from two different sources are
from the same phylogenetic tree. Recombination is thought to be the major reason
for causing such a discordance. However, there are other possibilities. A phylogenetic tree that represents the evolutionary pathways of a group of species is called a species tree, which can be regarded as the true tree. There are many methods to reconstruct this true tree. When it is constructed from one gene from each species, the inferred tree is a gene tree. The gene tree can be different from the species tree in two respects. First, the divergence of two genes sampled from two different species can predate the divergence of the two species. The second problem with the gene trees is that the topology of a gene tree may be different from that of the species. Therefore, more work has to be done to make the proposed method practically more useful in detecting recombination. We hope that this dissertation has achieved part of its goal as an initiator in this area.
BIBLIOGRAPHY


