ABSTRACT

TAO, WENLI. Testing Patterns of Nucleotide Substitution Rates at Multiple Genes. (Under the direction of Spencer V. Muse.)

Studying patterns of nucleotide substitution rates at multiple loci can help provide clues to the evolution and function of genes. The computational drawback of the maximum likelihood version of relative ratio tests becomes a concern when a large number of pairwise comparisons are performed among multiple genes. We propose a new version of relative ratio test, including four procedures, based on the use of pairwise sequence distances. The first is based on ANOVA two-way model and allows covariances between branch lengths. The second method applies generalized estimation equations (GEEs) to Poisson regression in a log-linear model. The third one is a nonparametric approach based on bootstrap percentile confidence intervals. The fourth method is based on weighted least squares estimation with covariances.

Simulation studies have been conducted to compare Type I errors and powers between the likelihood version of relative ratio tests and the first three proposed methods. The formulas have been derived for the last method as well as the numerical steps. The ANOVA-based method is the least computationally expensive and it has desirable Type I errors in most cases as well as good powers. The bootstrap-based method is the slowest among the four methods, but with smallest Type I errors and powers similar to the ANOVA-based method. The Likelihood-based method is the second slowest and has more desirable Type I errors than those of the ANOVA-based method, but has less powers than the ANOVA-based method. The GEE-based method is suitable only for very long genes, but has good statistical properties.

The ANOVA-based method is applied to mtDNA sequences from a broad range of animal mitochondrial genomes. The results indicate that it is not uncommon that branch lengths are conserved well among animal mitochondrial genes.

Key words: ANOVA with covariances, GEEs, Poisson regression in a log-linear model, bootstrap, weighted least squares with covariances, mitochondrial genome.

TESTING PATTERNS OF NUCLEOTIDE SUBSTITUTION RATES AT MULTIPLE GENES

by

Wenli Tao

A dissertation submitted to the Graduate Faculty of

North Carolina State University

in partial fulfillment of the

requirements for the Degree of

Doctor of Philosophy

BIOMATHEMATICS

Raleigh

2002

APPROVED BY:

Bruce S. Weir	Jeffrey L. Thorne
Edward S. Buckler	Spencer V. Muse CHAIR OF ADVISORY COMMITTEE

DEDICATION

To my parents and husband

BIOGRAPHY

I was born in Wuhan, China on June 10, 1968 to Yunkui Tao and Jiaoqin Zhang. I received my high school diploma from The Second High School of Huangshi in Hubei, China in 1986, and in the same year entered Tongji Medical University to study medicine. I graduated in 1991 with a Medical Degree, and went to Hubei Maternal and Children Hospital as a resident in pediatrics in July of 1991. During the time I also was involved in several clinical and epidemiological projects sponsored by WHO and UNICEF, where I needed to work with data. This allowed me to recognize the importance of statistics in research.

In August of 1996 I entered the Biomathematics Graduate Program at North Carolina State University in Raleigh. In August of 1997 I received a Master of Biomathematics degree with a minor in Statistics, and decided to pursue a doctorate degree in Biomathematics. In the summer and fall of 1998 I worked full-time at SAS Institute as an intern, where I had been involved in genetic data analysis. I found that I was very interested in the area of statistical genetics. In the spring of 1999 I returned to school to continue my study. My biggest blessings came on May 20, 1999, and September 13, 2000, when my daughters Katherine and Angie were born. With two little babies in hand, I once struggled to decide which direction I should go. Two persons helped me circumvent myself to make the right decision. My husband Neal dismissed my concerns about hardships and encouraged me to continue to pursue my Ph.D. My advisor, Spencer Muse, believed in my abilities and continued to support me financially.

My future plan is to work in the pharmaceutical/biotech industry in which I would use my multidisciplinary education to help bring good medicines to life.

ACKNOWLEDGMENTS

The list of people who contributed directly or indirectly to this dissertation is large. The first should go to Dr. Spencer Muse for his guidance and continuous encouragement, particularly, for his unfailing belief in my abilities and his invaluable financial support. I would like to express my appreciation to Dr. Bruce Weir for simply being who he is, a man of high intelligence and integrity. I also thank him for opening my eyes to the field of statistical genetics. I thank Dr. Jeffrey Thorne for his interesting insights and the challenge to think, and the thoroughness of his reviews, which greatly help me understand the theory of molecular evolution and enhance the quality of the content. I also want to thank Dr. Edward Buckler for his friendly manner, and providing me the opportunity to participate the process of microarray experiments.

Special thanks go to Dr. Zhao-Bang Zeng, Mr. Gaorui Wei, Dr. Yin Yang and Mr. Wayne Mitchell, who helped me step out of the hardship in my family during the spring of 1999. Without their gracious help, my family wouldn't end up such a happy way, and I wouldn't be able to successfully finish my degree. I would like to thank Ms Ann Ethridge for her help in many administrative details required to complete my degree, most importantly thanks for her encouragement and friendship. I want to thank Ms Debbie Hibbard, Ms Amy Elkins, and Ms Lisa Barefoot for their generous help and friendly smiles.

I thank my friends Xiuying Wang, Qifu Zheng, and their daughter Meiyi for their help and support. A special thanks to Xiuying for her help with Katherine so that I could focus on my thesis in the last month.

I thank my husband Neal for all of his love and faith in my abilities. Without his support and encouragement, I would never be able to go so far. I thank my parents in China, who have missed me greatly, for their continuous support and help all throughout the years of my education. The final thanks go to my daughters Katherine and Angie for being so special and bringing us so much joy.

Table of Contents

List of Tables

List of Figures

Chapter 1

Introduction

The study of variations and patterns of nucleotide substitution rates has drawn a great deal of attention in the last two decades with the explosion of DNA sequence data available to researchers. Indeed, knowledge of the patterns of nucleotide substitution is essential both to our understanding of molecular sequence evolution and to reliable estimation of phylogenetic relationships. In addition to testing the central issue in debates over the mode of molecular evolution, studying the rates of change in DNA sequences can provide insights into gene functions. For example, it has long been known that amino acids in structurally important protein regions tend to evolve more slowly than amino acids in less important regions (Kimura and Ohta 1974). Similar arguments apply to promoter regions embedded in non-coding DNA; in some cases, promoter regions have been identified because of their slow rate of evolution (Zurawski *et al.* 1987). With the advent of genomic technologies, it is common to obtain DNA sequence data for a set of species at multiple genes. The challenge will be to elucidate the functions of these genes. To this end, a comparative evolutionary approach will prove important. Therefore it is interesting to compare patterns of substitution rates on multiple genes simultaneously.

This project was motivated by work of Muse and Gaut (1997) on studying patterns of nucleotide substitution rates at multiple genes. In their work Muse and Gaut developed a statistical approach based on maximum likelihood method for simultaneous analysis of two or more loci. The test they proposed was termed as the relative ratio test, and it examined whether or not nucleotide substitution rates across evolutionary lineages had the same

relative magnitudes at two loci. Although the maximum likelihood method has recently become a major approach for sequence data analysis, its computational drawback can still be a concern if one wants to perform pairwise comparisons among multiple genes of interest. For example, if we are interested in performing pairwise comparisons among 500 genes in human genome, there will be total $\binom{500}{2} = 124750$ tests. If each test takes 1 minute, it needs about 87 days to finish this task if we only have one processor available. It is necessary to have an alternative method to reduce such computational burden. In addition, Muse and Gaut (1997) did not study the statistic properties of the maximum likelihood version of relative ratio tests. I therefore developed a new version of the relative ratio test based on the use of pairwise sequence distances, and compared the power of the likelihood-based method with the newly proposed methods. Several procedures are proposed to examine the same biological question as that of Muse and Gaut (1997).

In the introduction, I try to give an overview of the biological background and the basic concepts of the relative ratio tests in the terms of maximum likelihood and pairwise distance methods.

1.1 Nucleotide Sequences

1.1.1 Introduction of DNA Sequences

In most organisms, genetic information is carried by DNA (deoxyribonucleic acid) molecules. DNA is composed of two complementary chains twisted around each other to form a helix. Each chain is a long string of the four nucleotides or bases: two purines, adenine (**A**) and guanine (**G**), and two pyrimidines, cytosine (**C**) and thymine (**T**). The **A**'s complement with **T**'s, and so do **C**'s and **G**'s. DNA sequences are commonly represented by a sequence of letters such as ACGTTTGCCTTACAGA, with each letter corresponding to one of nucleotide bases. Usually the unit for DNA sequences is a base pair (*bp*).

A gene is an ordered sequence of nucleotides located in a particular position on a particular chromosome that encodes a specific functional product of either a protein or RNA molecule. Genes are a small portion of DNA sequences in some organisms. A locus usually

represents the position on a chromosome of a gene or the DNA at that position. The usage of terms gene and locus are sometimes interchangeable.

The earth's present-day species have developed from earlier, different species by evolutionary processes. The degree of relatedness between organisms or species can be estimated from the similarity of their DNA sequences. Although the processes of DNA replication from parents to offspring are highly accurate, mutations occur at a small but non-negligible rate that results in small differences between parents and offspring. Mutations result in the variability of organisms within a species which increases the chance that some organisms can survive in the face of large changes in the environment. Such small differences can accumulate in successive generations with roughly steady state. This makes it possible to understand the history of genes and species via comparative study of DNA sequences.

Mutation is any heritable change in DNA sequences. There are several mutational processes, including substitutions, recombination, deletions, insertions, and inversions. It is difficult to estimate the amount of mutation directly. Usually, the amount of substitution is used to reflect the degree of mutation. Nucleotide substitution is a basic process in the evolution of DNA sequences in which one nucleotide is replaced by another. In addition, substitution is relatively easy to study. Substitutions can be classified in two ways, one of which is based on resulting changes of coding in amino acid. The second way is to divide substitutions into transition and transversion. The exchanges within the purines or the pyrimidines (*i.e.*, $\mathbf{A} \leftrightarrow \mathbf{G}$ and $\mathbf{C} \leftrightarrow \mathbf{T}$) are called transition, while transversions are substitutions between the purines and pyrimidines (*e.g.*, $\mathbf{A} \leftrightarrow \mathbf{C}$ and $\mathbf{G} \leftrightarrow \mathbf{T}$). Accordingly, there are total four types of transitions and eight types of transversions. In general, the transition events happen more frequently than those of transversion. On the other hand, in protein coding regions two types of substitutions are identified based on the coding changes: synonymous substitutions and nonsynonymous substitutions. Synonymous substitutions are changes that do not result in a new amino acid, while nonsynonymous substitutions result in an amino acid change. Some authors prefer to naming them as silent and replacement

substitutions, respectively. Generally, synonymous substitutions occur at the third position of codons, while nonsynonymous substitutions mainly occur at the first or the second positions of codons. For example, the codon CCU (Pro) can have three synonymous substitutions, to CCC (Pro), CCA (Pro), or CCG (Pro), and six nonsynonymous substitutions, to UCU (Ser), ACU (Thr), GCU (Ala), CUU (Leu), CAU (His), or CGU (Arg).

1.1.2 Molecular Clocks

The molecular clock hypothesis asserts that the rate of evolution at the molecular level is approximately constant over time in all evolutionary lineages (Zuckerkandl and Pauling 1965). The concept of a molecular clock has had a great impact on the study of evolution, therefore debate on it has stimulated a number of empirical and theoretical projects. Observations of variation in nucleotide substitution rates in a variety of organisms have suggested that time-calibrated molecular clock is not universal at DNA level (*e.g.*, Li *et al.* 1992; Wolfe *et al.* 1987; Muse and Gaut 1994; Moran 1996). According to the neutral theory (Kimura 1968), substitution rates are expected to be inversely proportional to generation time. In other words, organisms with long generation times may have slower rates of nucleotide substitution than organisms with short generation times. The rationale behind this prediction is that if the number of germ-line cell divisions is equivalent between two organisms, then the organism with a shorter generation time has more germ-line cell divisions per unit time, which, in turn, results in a higher nucleotide substitution rate per unit time (Wu and Li 1985). Although some empirical studies supported the generation-time effect, there is still a debate as to the generality of a generation-time calibrated or time-calibrated clock in mammalian lineages (reviewed by Easteal *et al.* 1995). Except for generation-time effect, many other factors have been considered to contribute to the variation among substitution rates. For instance, metabolic rate has been postulated to affect mutation rates (Martin and Palumbi 1993). High metabolic rates may cause high concentrations of DNA altering free oxygen radicals and then affect mutation rates. It has also been suggested that phylogenetic groups with fast speciation rates will have fast substitution rates (Bousquet *et al.* 1992). Fidelity of DNA polymerase has also been hypothesized to influence

nucleotide substitution rates (Wu and Li 1985; Britten 1986). Authors have suggested that substitution rates also depend on population structure and the strength and direction of selection. For example, rates of substitution for mildly deleterious mutations are greater in small populations (Morgan 1996). Many forms of selection, say codon bias, are expected to be locus specific, but some selection will affect multiple loci (Ohta 1992). These molecular clock concepts make explicit predictions about correlations between nucleotide substitution rates and either time or life history traits, and the challenge is how to characterize the rates and patterns of nucleotide substitutions.

1.1.3 Nucleotide Substitution Rate

The change in nucleotides over time is a basic process in the evolution of DNA sequences. It is critical to study such changes in nucleotide sequences in order to estimate evolutionary rates, to estimate divergence times, and to reconstruct evolutionary trees. Nucleotide substitutions are commonly detected by comparing differences of nucleotides between two alignments of homologous DNA sequences. The number of nucleotide substitutions between two sequences (Figure 1.1) is generally expressed in terms of the number of substitutions per nucleotide site (K) instead of the total number (N) of substitutions between the two sequences. If we know L - the number of nucleotide substitution sites compared between the two sequences and N, then we can have $K = N/L$, which is referred as the evolutionary distance between two homologous sequences. For example, in Figure 1.1, Sequence 1 and Sequence 2 are two homologous sequences diverged from the same ancestral sequence ACCTAGAG T years ago. The length of the sequence is $L = 8$. There is a total of four nucleotide substitutions between the sequences since they shared the common ancestor (i.e., $N = 4$). The number of substitutions per site is then calculated by $K = 4/8 = .5$. In fact, the true history of evolution is unknowable in practice. We can only observe three differences between the two sequences at positions $1, 5$, and 6 due to multiple substitutions at the first position. That is, the number of observed differences between two sequences is always an underestimate of the true number of evolutionary changes that took place. Since multiple

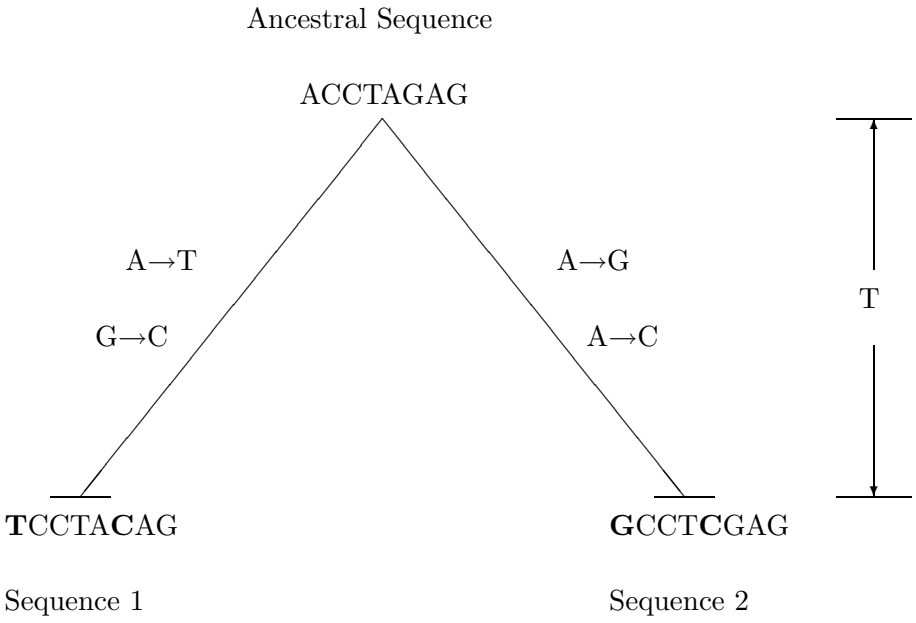


Figure 1.1: Divergence of two homologous sequences from a common ancestral sequence T years ago. The bold letters represent substitution sites

substitutions could occur at any site and it is impossible to observe such replacements, usually it is necessary to correct for multiple hits before estimating substitution rates. Without such a correction, the rate of change in sequences will always be underestimated. In order to count such multiple substitutions, mathematical models have been developed, and they will be described in the following section.

We need to estimate the nucleotide substitution rate for different lineages. The rate of nucleotide substitution is defined as the number of substitutions per site per year. There are two basic measurements of nucleotide substitution rates: "absolute" rate of substitution and "relative" rate of substitution. In order to estimate the absolute rates, we need to know the homologous nucleotide sequences from two taxa and the divergence time between the taxa. Let λ be the absolute rate of nucleotide substitution per site per year. Given the estimated number of nucleotide substitution per site, $\hat{K}$, and the estimated divergence time,

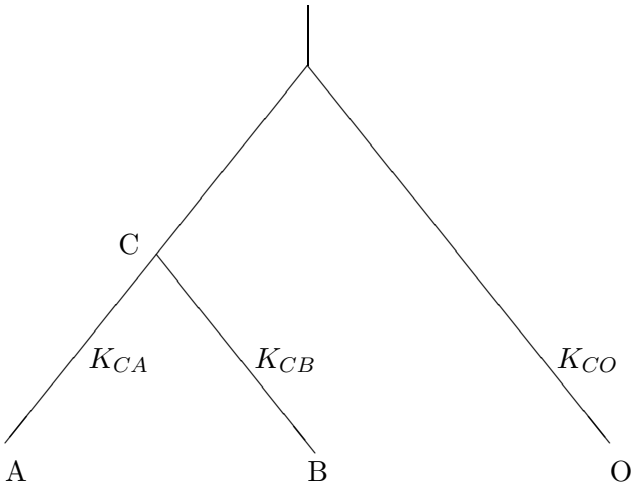


Figure 1.2: The rooted tree of the species A, B, O, assuming that O is the known outgroup. C denotes the common ancestor of species A and B

$\hat{T}$, the absolute rate of nucleotide substitution per site per year can be estimated as:

$$\hat{\lambda} = \hat{K}/2\hat{T}.$$

That is, $\hat{K} = 2\hat{\lambda}\hat{T}$. K can be estimated via a variety of nucleotide substitution models which will be reviewed later. The divergence time T is usually inferred from fossil data. Traditionally, estimation of absolute rates requires the assumption that rates of evolution are equal in the two lineages since they diverged, i.e., an assumption of time-calibrated molecular clock. In reality, unavailability and unreliability of fossil data often make it impossible to directly estimate the absolute rate of substitution under the assumption of time-calibrated molecular clock. To solve this problem, several approaches for estimating absolute rates recently have been proposed by relaxing the assumption of constant rate among lineages (Sanderson 1997 & 2002, Thorne *et al.* 1998, Huelsenbeck *et al.* 2000, Kishino *et al.* 2001). Although those approaches used different statistical methods, they shared the same idea of estimating divergence times along with absolute rates. Sanderson (1997) proposed a nonparametric rate smoothing method in which absolute rates and divergence times were

estimated by a least-squares smoothing method. Thorne *et al.* (1998) proposed a hierarchical model of rate evolution in which the logarithm of the rate on the same branch was assumed to be a normal distribution. The Bayesian approach was applied to estimate the rates and times. Huelsenbeck *et al.* (2000) followed the idea of Thorne *et al.* (1998), but they allowed rates to vary across lineages according to a compound Poisson process in which the amount of rate changes was assumed to be a gamma distribution. To take advantages of both nonparametric and parametric methods, Sanderson (2002) proposed a semiparametric smoothing method to estimate the rates and times by using penalized likelihood.

Relative substitution rate has been commonly used to compare the difference of substitution rates, either between genes or between lineages, without knowledge of divergence time. Figure 1.2 shows the idea of obtaining the relative rate. This approach requires at least three homologous DNA sequences in which one serves as an outgroup and the remaining sequences are ingroups. The purpose of the outgroup is to root the tree. Let K_{ij}, or K_{ji}, be the number of substitutions per site separating nodes i and j. It is obvious that $K_{AO} = K_{AC}+K_{CO}$ and $K_{BO} = K_{BC}+K_{CO}$ and by subtraction, $K_{AO}-K_{BO} = K_{AC}-K_{BC}$. Therefore comparison between K_{AO} and K_{BO} is equivalent to comparison between K_{AC} and K_{BC}. In other words, using an outgroup makes it possible to estimate the number of substitution events on the branches leading from the common ancestor C to ingroups A and B, respectively. Such estimation is not possible using only the ingroup sequences A and B. It follows that the relative rate of substitution can be obtained by the estimated $\hat{K}_{AC}$ and $\hat{K}_{BC}$ as

$$\hat{R} = \hat{K}_{AC}/\hat{K}_{BC}.$$

Note that the relative rate R is independent of time dimension since divergence time will be canceled out when ratio is taken.

Table 1.1: Common nucleotide substitution models

Models	Constraint	Comment	Reference
JC69	$\theta_1 = \theta_2 = \theta_3 = \theta_4 = \theta_5 = \theta_6$, $\pi_A = \pi_C = \pi_G = \pi_T$	all substitutions equally probable; equal base frequencies.	Juke and Cantor(1969)
F81	$\theta_1 = \theta_2 = \theta_3 = \theta_4 = \theta_5 = \theta_6$	all substitutions equally probable; base frequencies may be unequal	Felsenstein (1981)
K80	$\theta_1 = \theta_3 = \theta_4 = \theta_6$; $\theta_2 = \theta_5$, $\pi_A = \pi_C = \pi_G = \pi_T$	separate transition and transversion rates; equal base frequencies	Kimura (1980)
HKY85	$\theta_1 = \theta_3 = \theta_4 = \theta_6$; $\theta_2 = \theta_5$.	separate transition and transversion rates; base frequencies may be unequal.	Hasegawa et al. (1985)
TN93	$\theta_1 = \theta_3 = \theta_4 = \theta_6$	two different transition rates, and one transversion rate; base frequencies may be unequal.	Tamura and Nei (1993)
REV	none	all substitutions occur at different rates; base frequencies may be unequal.	Tavare (1986)

1.1.4 Models of Nucleotide Substitution

It was recognized long ago that nucleotide substitution could be described using a stationary homogeneous Markov process (Yang 1994), which was characterized by a rate matrix whose elements represent instantaneous substitution rates among the four nucleotides. Different restrictions on the matrix result in various mathematical models of substitution that are summarized in Table 1.1 (Muse 2000). Almost all the models proposed in the literature are a special form of the General Reversible Process model (REV) (Dayhoff *et al.* 1978, Tavaré

1986). The rate matrix for this model takes the form

$$\mathbf{Q}_{\mathrm{REV}} = \begin{array}{c} \\ A \\ C \\ G \\ T \end{array} \begin{array}{c} \begin{array}{cccc} \mathrm{A} & \mathrm{C} & \mathrm{G} & \mathrm{T} \end{array} \\ \begin{pmatrix} -\Sigma_1 & \theta_1\pi_C & \theta_2\pi_G & \theta_3\pi_T \\ \theta_1\pi_A & -\Sigma_2 & \theta_4\pi_G & \theta_5\pi_T \\ \theta_2\pi_A & \theta_4\pi_C & -\Sigma_3 & \theta_6\pi_T \\ \theta_3\pi_A & \theta_5\pi_C & \theta_6\pi_G & -\Sigma_4 \end{pmatrix} \end{array}.$$

The elements on the diagonal are the Σ_i $(i = 1, 2, 3, 4)$ which is the sum of all the other three elements on the ith row so the total sum on the row is zero. For example, at the first row $\Sigma_1 = \theta_1\pi_C + \theta_2\pi_G + \theta_3\pi_T$. The off-diagonal values in the matrix represent the instantaneous probabilities of changes from the nucleotides indexed by the row to those indexed by the column, $P_{ij}(\delta t)$. Note that every off-diagonal element is a product of two terms. The θ_i's are substitution parameters that control the relative rates of changes between different nucleotides. The π_i's are frequency parameters representing equilibrium base frequencies for the four nucleotides. The summation of the π_i's is one. The matrix satisfies the time reversibility condition of $\pi_i Q_{ij} = \pi_j Q_{ji}$. As Yang (1994) pointed out, it is unclear whether it is biologically reasonable to consider these two sets of parameters as representing different forces that affect nucleotide substitution, but this distinction is mathematically convenient. Although the precise probabilities of changes among different nucleotides may depend on the base frequencies, the general trends are governed by the substitution parameters.

Different combinations of substitution parameters and base frequencies will result in different nucleotide models. The model of Jukes and Cantor (1969) (JC69) is the simplest in that all the changes among the four nucleotides are assumed to occur with equal probability.

So the rate matrix of the JC69 model is defined as:

$$\mathbf{Q}_{\text{JC69}} = \begin{array}{c} \\ A \\ C \\ G \\ T \end{array} \begin{array}{c} \begin{array}{cccc} \text{A} & \text{C} & \text{G} & \text{T} \end{array} \\ \begin{pmatrix} -3\theta/4 & \theta/4 & \theta/4 & \theta/4 \\ \theta/4 & -3\theta/4 & \theta/4 & \theta/4 \\ \theta/4 & \theta/4 & -3\theta/4 & \theta/4 \\ \theta/4 & \theta/4 & \theta/4 & -3\theta/4 \end{pmatrix} \end{array}.$$

Note that all changes have the same substitution parameter θ and all nucleotides have the same base frequency, 1/4.

Kimura's model (1980) (K80) allows different rates of transition and transversions by equating $\theta_2 = \theta_5$ and $\theta_1 = \theta_3 = \theta_4 = \theta_6$. The rate of matrix of K80 model is given by:

$$\mathbf{Q}_{\text{K80}} = \begin{array}{c} \\ A \\ C \\ G \\ T \end{array} \begin{array}{c} \begin{array}{cccc} \text{A} & \text{C} & \text{G} & \text{T} \end{array} \\ \begin{pmatrix} -\Sigma_1 & \theta_1/4 & \theta_2/4 & \theta_1/4 \\ \theta_1/4 & -\Sigma_2 & \theta_1/4 & \theta_2/4 \\ \theta_2/4 & \theta_1/4 & -\Sigma_3 & \theta_1/4 \\ \theta_1/4 & \theta_2/4 & \theta_1/4 & -\Sigma_4 \end{pmatrix} \end{array}.$$

Felsenstein's model (1981) (Fel81) allows the four nucleotides to have unequal frequencies at equilibrium and have the same substitution parameter θ. The rate matrix of the Fel81 model is defined as:

$$\mathbf{Q}_{\text{Fel81}} = \begin{array}{c} \\ A \\ C \\ G \\ T \end{array} \begin{array}{c} \begin{array}{cccc} \text{A} & \text{C} & \text{G} & \text{T} \end{array} \\ \begin{pmatrix} -\Sigma_1 & \theta\pi_C & \theta\pi_G & \theta\pi_T \\ \theta\pi_A & -\Sigma_2 & \theta\pi_G & \theta\pi_T \\ \theta\pi_A & \theta\pi_C & -\Sigma_3 & \theta\pi_T \\ \theta\pi_A & \theta\pi_C & \theta\pi_G & -\Sigma_4 \end{pmatrix} \end{array}.$$

The model of Hasegawa *et al.* (1985) (HKY85) is the general case of the above three which allows both different rates for transition and transversions and different nucleotide

frequencies. The rate matrix of the HKY85 model is

$$\mathbf{Q}_{\text{HKY85}} = \begin{array}{c} \\ A \\ C \\ G \\ T \end{array} \begin{array}{c} \begin{array}{cccc} \text{A} & \text{C} & \text{G} & \text{T} \end{array} \\ \begin{pmatrix} -\Sigma_1 & \theta_2\pi_C & \theta_1\pi_G & \theta_2\pi_T \\ \theta_2\pi_A & -\Sigma_2 & \theta_2\pi_G & \theta_1\pi_T \\ \theta_1\pi_A & \theta_2\pi_C & -\Sigma_3 & \theta_2\pi_T \\ \theta_2\pi_A & \theta_1\pi_C & \theta_2\pi_G & -\Sigma_4 \end{pmatrix} \end{array}.$$

The model of Tamura and Nei (1993) (TN93) makes only the restriction that $\theta_1 = \theta_3 = \theta_4 = \theta_6$, allowing for two classes of transitions. It is easy to see that the model of HKY85 is a special case of the model of Tamura and Nei (1993). The rate matrix for the model TN93 is given by

$$\mathbf{Q}_{\text{TN93}} = \begin{array}{c} \\ A \\ C \\ G \\ T \end{array} \begin{array}{c} \begin{array}{cccc} \text{A} & \text{C} & \text{G} & \text{T} \end{array} \\ \begin{pmatrix} -\Sigma_1 & \theta_3\pi_C & \theta_1\pi_G & \theta_3\pi_T \\ \theta_3\pi_A & -\Sigma_2 & \theta_3\pi_G & \theta_2\pi_T \\ \theta_1\pi_A & \theta_3\pi_C & -\Sigma_3 & \theta_3\pi_T \\ \theta_3\pi_A & \theta_2\pi_C & \theta_3\pi_G & -\Sigma_4 \end{pmatrix} \end{array}.$$

1.1.5 Estimating Nucleotide Substitution Rate

Given the instantaneous rate matrix $\mathbf{Q}$, the substitution probability matrix for an interval of time t is given by

$$\mathbf{P}(t) = e^{\mathbf{Q}t}, \tag{1.1}$$

The matrix exponential $e^{\mathbf{Qt}}$ can be calculated by

$$e^{\mathbf{Q}t} = \sum_{i=1}^{4} e^{t\lambda_i}\mathbf{v}_i\mathbf{w}_i', \tag{1.2}$$

where λ_i, $i = 1, 2, 3, 4$, are the eigenvalues of matrix $\mathbf{Q}$, and $\mathbf{v}_i$ and $\mathbf{w}_i$ are the left and right eigenvectors related to the ith eigenvalue (i.e., $\mathbf{Q}\mathbf{v}_i = \lambda_i\mathbf{v}_i$ and $\mathbf{Q}'\mathbf{w}_i = \lambda_i\mathbf{w}_i$).

For the JC69 model, we have the eigenvalues and the corresponding eigenvectors for the

$\mathbf{Q}_{\text{JC69}}$ as follows,

$$\lambda_1 = 0, \quad \lambda_2 = \theta, \quad \lambda_3 = \lambda_4 = -2\theta.$$

$$\mathbf{v}_1 = \begin{pmatrix} 1 \\ 1 \\ 1 \\ 1 \end{pmatrix}, \quad \mathbf{v}_2 = \begin{pmatrix} 1/2 \\ 1/2 \\ -1/2 \\ -1/2 \end{pmatrix}, \quad \mathbf{v}_3 = \begin{pmatrix} 0 \\ 0 \\ 1/2 \\ -1/2 \end{pmatrix}, \quad \mathbf{v}_4 = \begin{pmatrix} 1/2 \\ -1/2 \\ 0 \\ 0 \end{pmatrix},$$

$$\mathbf{w}_1 = \begin{pmatrix} 1/4 \\ 1/4 \\ 1/4 \\ 1/4 \end{pmatrix}, \quad \mathbf{w}_2 = \begin{pmatrix} 1/8 \\ 1/8 \\ -1/8 \\ -1/8 \end{pmatrix}, \quad \mathbf{w}_3 = \begin{pmatrix} 0 \\ 0 \\ 1 \\ -1 \end{pmatrix}, \quad \mathbf{w}_4 = \begin{pmatrix} 1 \\ -1 \\ 0 \\ 0 \end{pmatrix}.$$

Using the equations (1.1) and (1.2), we have transitional probabilities for the model JC69

$$P_{ij}(t) = \begin{cases} \frac{1}{4} + \frac{3}{4}e^{-4\theta t/3} & i = j \\ \frac{1}{4} - \frac{1}{4}e^{-4\theta t/3} & i \neq j \end{cases}$$

Let p_t denote the probability that two homologous nucleotides are the same type at time t after diverging from their common ancestor. Note that

$$\begin{aligned} p_t &= \sum_j P_{ji}(t)P_{ij}(t) \\ &= \sum_j P_{ij}(t)P_{ji}(t) \\ &= P_{ii}(2t) \qquad \text{(by Chapman-Kolmogorov equation)} \\ &= \frac{1}{4} + \frac{3}{4}e^{-8\theta t/3}. \end{aligned}$$

Solving the above equation, the estimate of $K = 2\theta t$, the expected number of substitutions

separating two homologous sites, is given by

$$\hat{K} \;=\; \frac{3}{4}\ln\left(\frac{3}{4\hat{p}_t - 1}\right), \tag{1.3}$$

where $\hat{p}_t$ is the observed proportion of identical sites in the two sequences being compared. $\hat{K}$ is usually refereed as the evolutionary distance between the two sequences. It can be shown that $\hat{K}$ in (1.3) is a maximum−likelihood estimator of K.

For the HKY85 model, the eigenvalues of $\mathbf{Q}_{\text{HKY85}}$ are

$$\lambda_1 = 0, \quad \lambda_2 = -(\pi_Y\theta_2 + \pi_R\theta_1),$$

$$\lambda_3 = -\theta_2, \quad \lambda_4 = -(\pi_Y\theta_1 + \pi_R\theta_2),$$

and the corresponding eigenvectors are

$$\mathbf{v}_1 = \begin{pmatrix} \pi_T \\ \pi_C \\ \pi_A \\ \pi_G \end{pmatrix}, \quad \mathbf{v}_1 = \begin{pmatrix} 0 \\ 0 \\ 1 \\ -1 \end{pmatrix}, \quad \mathbf{v}_3 = \begin{pmatrix} \pi_R\pi_T \\ \pi_R\pi_C \\ \pi_Y\pi_A \\ \pi_Y\pi_G \end{pmatrix}, \quad \mathbf{v}_4 = \begin{pmatrix} 1 \\ -1 \\ 0 \\ 0 \end{pmatrix},$$

$$\mathbf{w}_1 = \begin{pmatrix} 1 \\ 1 \\ 1 \\ 1 \end{pmatrix}, \quad \mathbf{w}_2 = \begin{pmatrix} 0 \\ 0 \\ \pi_G/\pi_R \\ -\pi_A/\pi_R \end{pmatrix}, \quad \mathbf{w}_3 = \begin{pmatrix} 1/\pi_Y \\ 1/\pi_Y \\ -1/\pi_R \\ -1/\pi_R \end{pmatrix}, \quad \mathbf{w}_4 = \begin{pmatrix} \pi_C/\pi_Y \\ -\pi_T/\pi_Y \\ 0 \\ 0 \end{pmatrix},$$

where $\pi_Y = \pi_T + \pi_C$ and $\pi_R = \pi_A + \pi_G$. Applying these expressions into equations (1.1) and (1.2), we find the transition probabilities for the HKY85 model:

$$P_{ij}(t) \;=\; \begin{cases} \pi_j\left(1 + \frac{1-\Pi_j}{\Pi_j}e^{-\theta_2 t} - \frac{1}{\Pi_j}e^{-[\Pi_j\theta_1 + (1-\Pi_j)\theta_1]t}\right) & i \neq j, \text{ transition} \\ \pi_j\left(1 - e^{-\theta_2 t}\right) & i \neq j, \text{ transversion} \\ \pi_j\left(1 + \frac{1-\Pi_j}{\Pi_j}e^{-\theta_2 t}\right) + \frac{\Pi_j - \pi_j}{\Pi_j}e^{-[\Pi_j\theta_1 + (1-\Pi_j)\theta_2]t} & i = j, \text{ no change,} \end{cases}$$

where Π_j is the frequency of either purines or pyrimidines, depending on the classification of the "target" nucleotide, j. For example, when $i = A$, and $j = G$, there is a transition to a purine and $\Pi_G = \pi_A + \pi_G$. When $i = A$, and $j = C$, there is a transversion to a pyrimidine and $\Pi_C = \pi_C + \pi_T$.

Note that the substitution probabilities in both models depend only on the product of θ and t, which reflects the confounding of these two parameters. Thus the estimates of substitution rates along evolutionary lineages are usually estimated by branch lengths, which may be better interpreted as an average, over time, of the amount of sequence change per site.

There are some mathematical requirements for the substitution model in order to obtain an analytical formula for estimating the distance separating two sequences (Yang 1994). Only do the JC69, K80, F81 and TN93 meet the requirements among the models listed above. It is impossible to derive the analytical formula for estimating the expected number of nucleotide substitutions for the HKY85 model. Note in the model of TN93, parameters θ_1, θ_2, and θ_3 stand for the rates of transitional replacement between purines, between pyrimidines, and of transversional change, respectively. If we let $\theta_1 = \theta_2$ in the model of TN93, it becomes the HKY85 model. Thus, the formula of evolutionary distance from the TN93 model may be used to get the estimate of distance for HKY85. When the nucleotide frequencies were in equilibrium, Tamura and Nei (1993) derived the average rate of nucleotide substitution per site as

$$\lambda = 2\pi_A\pi_G\theta_1 + 2\pi_T\pi_C\theta_2 + 2\pi_R\pi_Y\theta_3,$$

where $\pi_R = \pi_A + \pi_G$ and $\pi_Y = \pi_C + \pi_T$ as defined before. The expected number of nucleotide substitutions between two homologous sequences diverged t evolutional time unites (years or generations) ago is given by $K = 2\lambda t$ where

$$K = 4\pi_A\pi_G\theta_1 t + 4\pi_T\pi_C\theta_2 t + 4\pi_R\pi_Y\theta_3 t.$$

Denote the expected proportions of nucleotide sites showing transitional differences between purines as P_1 and between pyrimidines as P_2 and of those showing transversional differences as S. Tamura and Nei (1993) derived the following formulas

$$P_1 = \frac{2\pi_A\pi_G}{\pi_R}\pi_R + \pi_Y \exp(-2\theta_3 t) - \exp[-2(\pi_R\theta_1 + \pi_Y\theta_3)t], \tag{1.4}$$

$$P_2 = \frac{2\pi_T\pi_C}{\pi_Y}\pi_Y + \pi_R \exp(-2\theta_3 t) - \exp[-2(\pi_Y\theta_2 + \pi_R\theta_3)t], \tag{1.5}$$

$$S = 2\pi_R\pi_Y[1 - \exp(-2\theta_3 t)]. \tag{1.6}$$

In the model of HKY85, the rates of transition are equal, $\theta_1 = \theta_2 = \delta$. Consequently, the expected proportions of nucleotide sites showing transitional differences should be the sum of differences between purines and differences between pyrimidines, $P = P_1 + P_2$. For consistency, let $\theta_3 = \rho$ denote the transversional rate. The expected number of nucleotide substitution between two sequences becomes

$$K = 4\delta(\pi_A\pi_G + \pi_T\pi_C)t + 4\pi_R\pi_Y\rho t.$$

In practice, P and S are estimable from sequence comparison and the four base frequencies can be estimated by the average nucleotide frequencies of the two sequences compared. To get the estimate of K, we should get the estimates of both δ and ρ. From (1.6), we can easily obtain the formula for the estimate of ρt as

$$\widehat{\rho t} = -\frac{1}{2}\ln\left(1 - \frac{S}{2\pi_R\pi_Y}\right).$$

By putting P_1 and P_2 together, we have

$$\begin{aligned} P = {} & 2\pi_A\pi_G + 2\pi_T\pi_C + \left(\frac{2\pi_A\pi_G\pi_Y}{\pi_R} + \frac{\pi_T\pi_C\pi_R}{\pi_Y}\right)\exp(-2\rho t) \\ & -\frac{2\pi_A\pi_G}{\pi_R}\{\exp[-2(\pi_R\delta + \pi_Y\rho)t]\} \\ & -\frac{2\pi_T\pi_C}{\pi_Y}\{\exp[-2(\pi_Y\delta + \pi_R\rho)t]\}. \end{aligned} \tag{1.7}$$

Although it is impossible to get an analytical formula for δt from equation (1.7), we still can obtain the estimate of δt by taking the advantage of the computer. Consider the function

$$\begin{aligned} f(\delta) &= 2\pi_A\pi_G + 2\pi_T\pi_C + \left(\frac{2\pi_A\pi_G\pi_Y}{\pi_R} + \frac{\pi_T\pi_C\pi_R}{\pi_Y}\right)\exp(-2\rho t) \\ &\quad -\frac{2\pi_A\pi_G}{\pi_R}\{\exp[-2(\pi_R\delta + \pi_Y\rho)t]\} \\ &\quad -\frac{2\pi_T\pi_C}{\pi_Y}\{\exp[-2(\pi_Y\delta + \pi_R\rho)t]\} - P. \end{aligned}$$

Assume that $f(\delta) = 0$. By taking the first derivative with respect to δ, we obtain

$$\begin{aligned} f'(\delta) &= 4\pi_A\pi_G\{\exp[-2(\pi_R\delta + \pi_Y\rho)t]\} + 4\pi_T\pi_C\{\exp[-2(\pi_Y\delta + \pi_R\rho)t]\} \\ &> 0. \end{aligned}$$

Therefore the function $f(\delta)$ is a strictly increasing function on $(-\infty, +\infty)$. Since $f(\delta) = 0$, $\lim_{\delta\to+\infty} f(\delta) > 0$, and $\lim_{\delta\to-\infty} f(\delta) < 0$, there exist a positive integer N_1 and a negative integer N_2 such that

$$f(N_1) > 0 \quad (N_1 > 1),$$

and

$$f(N_2) < 0 \quad (N_2 < -1),$$

which satisfies the condition for the bisection method for root finding (Press *et al.* 1992). Accordingly, the estimate of δt can be obtained by applying the bisection method to solve the function $f(\delta) = 0$.

In general, the distance estimate for any model can be found using numerical methods, even when analytical formulas are not available.

1.2 Comparison of Patterns of Substitution Rates at One Loci

1.2.1 Relative Rate Test

Debate on the molecular clock hypothesis is always concerned about the divergence time. In general, it is impossible to compare evolutionary rates among species, and the estimated substitution rates can be used for this purpose. Unavailable and unreliable fossil data made it difficult to estimate the actual substitution rates and test the hypothesis. To solve this problem, the relative rate test was proposed by Sarich and Wilson (1973). As mentioned in the previous section, by introducing an outgroup species the branch lengths along corresponding lineages can be estimated without knowledge of divergence time. The prior knowledge needed to know is that the outgroup species should be more relatively distanced than either of two species being compared. In order to know whether or not substitution rates are same between two species A and B since they diverged from the common ancestor C, the reference species O is used to root the tree as shown in Figure 1.2. As usual, letting K_{ij} be the expected number of substitutions per site separating species i and j, we have

$$K_{AO} = K_{AC} + K_{CO}, \tag{1.7a}$$

$$K_{BO} = K_{BC} + K_{CO}, \tag{1.7b}$$

$$K_{AB} = K_{AC} + K_{BC}, \tag{1.7c}$$

where K_{AO}, K_{BO}, and K_{CO} can be directly estimated from the nucleotide sequences based on the nucleotide models described in the previous section. Solving these equations, we can obtain the estimate of K_{AC}, K_{BC}, and K_{CO}:

$$\begin{aligned} K_{AC} &= \frac{1}{2}(K_{AO} + K_{AB} - K_{BO}), \\ K_{BC} &= \frac{1}{2}(K_{AB} + K_{BO} - K_{AO}), \\ K_{CO} &= \frac{1}{2}(K_{AO} + K_{BO} - K_{AB}). \end{aligned}$$

That is, we can estimate the number of substitution per site on the branches leading from the common ancestor C to ingroups A and B by using the outgroup O. The concept of relative rate of substitution provides a sufficient way to test the molecular clock hypothesis without knowing the divergence time. In order to well understand the concept of relative rate test, essential to the contents of the later chapters, I will briefly introduce two different versions of relative rate test.

1.2.2 Distance-based Methods

The evolutionary distance between a pair of sequences is typically referred as the estimate of average number of substitutions per site. In other words, it is the sum of the branch lengths separating two species being compared. There are several distance-based relative rate tests (*e.g.,* Wu and Li 1985, Li *et al.* 1985 and Li and Bouquet 1992). The basic idea is to test the difference of branch lengths between two species of interest by constructing a form of Z-test. Referring back to Figure 1.2, we want to test the equality of rates between species A and species B. Because the time that passed since species A and B diverged from the last common ancestral C is equal for both lineages, K_{AC} and K_{BC} should be equal according to the molecular clock hypothesis. From equation (1.7), it is easy to obtain $K_{AC} - K_{BC} = K_{AO} - K_{BO}$. Thus, testing $H_0 : K_{AC} - K_{BC} = 0$ is equivalent to testing the null hypothesis $H_0 : K_{AO} - K_{BO} = 0$. The Z-test is given by

$$Z = \frac{\hat{K}_{AO} - \hat{K}_{BO}}{\sqrt{\mathrm{V}(\hat{K}_{AO} - \hat{K}_{BO})}},$$

where

$$\mathrm{V}(\hat{K}_{AO} - \hat{K}_{BO}) = \mathrm{V}(\hat{K}_{AO}) + \mathrm{V}(\hat{K}_{BO}) - 2\mathrm{Cov}(\hat{K}_{AO}, \hat{K}_{BO}).$$

Note that $\hat{K}_{AO} = \hat{K}_{AC} + \hat{K}_{CO}$ and $\hat{K}_{BO} = \hat{K}_{BC} + \hat{K}_{CO}$ and that $\hat{K}_{AC}$ and $\hat{K}_{BC}$ are independent of each other since lineages A and B evolve independently. Thus, $\mathrm{Cov}(\hat{K}_{AO}, \hat{K}_{BO}) = \mathrm{Var}(\hat{K}_{CO})$ (Nei *et al.* 1985). Consequently, we have

$$\mathrm{V}(\hat{K}_{AO} - \hat{K}_{BO}) = \mathrm{V}(\hat{K}_{AO}) + \mathrm{V}(\hat{K}_{BO}) - 2\mathrm{Var}(\hat{K}_{CO}).$$

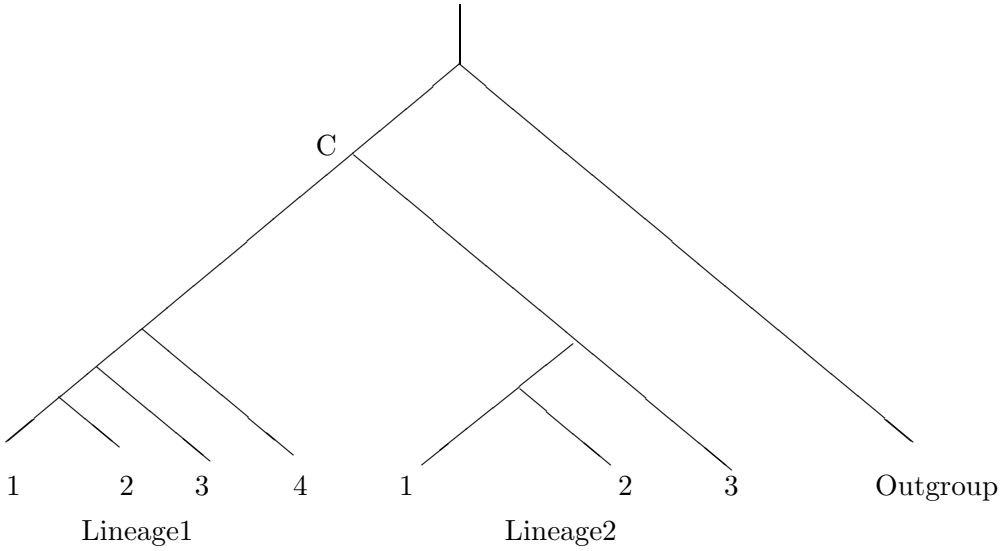


Figure 1.3: The phylogenetic tree of the test of Li and Bouquet (1992): the tree consists of three lineages, with Lineage 1 having $n = 4$ taxa, with Lineage 2 having $m = 3$ taxa and an outgroup. C is the common ancestor of Lineage 1 and Lineage 2

The Z-test is approximately distributed as $N(0, 1)$.

The tests of Wu and Li (1985) and Li *et al.* (1985) only compared two taxa simultaneously, while Li and Bouquet (1992) extended them to cases where more than one taxon are sampled for one nucleotide sequence in two lineages (Figure 1.3). Suppose that a tree consists of two lineages, with Lineage 1 having n taxa and with Lineage 2 having m taxa. Define $N_i^{(1)}$ as the number of nucleotides compared between the ith sequence in Lineage 1 and the outgroup sequence, $i = 1, \cdots, n$, and define $N_j^{(2)}$ as the number of nucleotides compared between the jth sequence in Lineage 2 and the outgroup sequence, $j = 1, \cdots, m$. Let x and y be the sums of $N_i^{(1)}$ and $N_j^{(2)}$, respectively. Define $K_i^{(1)}$ and $K_j^{(2)}$ to be the number of nucleotide substitutions between the ith sequence in Lineage 1 and the outgroup, and jth sequence in Lineage 2 and the outgroup, respectively. The average numbers of nucleotide substitutions between Lineage 1 and the outgroup, and between Lineage 2 and the

outgroup are given by

$$\begin{aligned} K_1 &= \sum_{i=1}^{n}(N_i^{(1)}K_i^{(1)})/x, \\ K_2 &= \sum_{i=1}^{m}(N_j^{(2)}K_j^{(2)})/y. \end{aligned}$$

Then the form of Z-test can be constructed as before,

$$Z = \frac{K_1 - K_2}{\sqrt{\mathrm{V}(K_1 - K_2)}},$$

which is also approximately distributed as $N(0, 1)$.

1.2.3 Maximum Likelihood-based Method

Felsenstein (1981) proposed to use maximum likelihood methods to estimate the phylogenetic trees. Following the idea of Felsenstein (1981), a more powerful and flexible likelihood ratio version of relative rate test was formally developed by Muse and Weir (1992). The basic idea of this version of relative ratio test is the same as the distance-based methods. To understand the test, first we should know how to form the likelihood function for phylogenetic tree (Felsenstein 1981).

Suppose we have three species A, B, and O and know the common ancestor D for all and the common ancestor C only for A and B. Denote the nucleotide at site k of sequence i as S_{ik} ($i = A, B, C, D$, and O). Following Felsenstein (1981) the likelihood function for site k is given by

$$\begin{aligned} l_k = \sum_{S_{Dk}=1}^{4}\sum_{S_{Ck}=1}^{4} & \pi_{S_{Dk}} P_{S_{Dk},S_{Ck}}(K_{CD}) P_{S_{Dk},S_{Ok}}(K_{DO}) \\ & P_{S_{Ck},S_{Ak}}(K_{AC}) P_{S_{Ck},S_{Bk}}(K_{BC}) \end{aligned}$$

where the term $P_{S_{ik},S_{jk}}(K_{ij})$ ($j = A, B, C, D$, and O) represents the probability that a site being in state S_{jk} was initially in state S_{ik} (either A, C, G, or T) after period of time, while $K_{ij} = \mu_{ij}t_{ij}$, is the product of substitution rate and time for the specific branch in the tree,

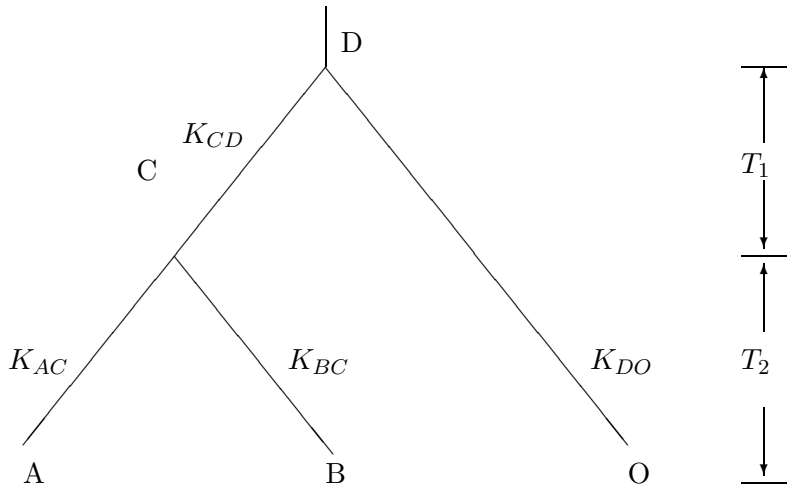


Figure 1.4: The maximum likelihood-based relative rate test: three species A, B, and O and know the common ancestor D for all and the common ancestor C only for A and B

i.e. branch length. Substitution rates confound with divergence time, so it is impossible to estimate substitution rate directly without knowledge of time information. Felsenstein (1981) provided the "pulley principle" to reduce the computational demand. That is, for time-reversible models the likelihood function is the same for all rooted trees corresponding to a single unrooted tree (Figure 1.5). Therefore the previous equation can be simplified by

$$l_k = \sum_{S_{Ck}=1}^{4} \pi_{S_{Ck}} P_{S_{Ck},S_{Ok}}(K_{CO}) P_{S_{Ck},S_{Ak}}(K_{AC}) P_{S_{Ck},S_{Bk}}(K_{BC}),$$

which reduces the two summations to one and four branches to three.

Assuming that substitution at one position in a sequence occurs independently from events at other positions, the overall likelihood function for n sites is the product of the n individual site likelihoods,

$$L = \prod_{k=1}^{n} l_k.$$

Suppose we wish to test the same equality as the distance-based methods: $H_0 : K_{AC} = K_{BC}$,

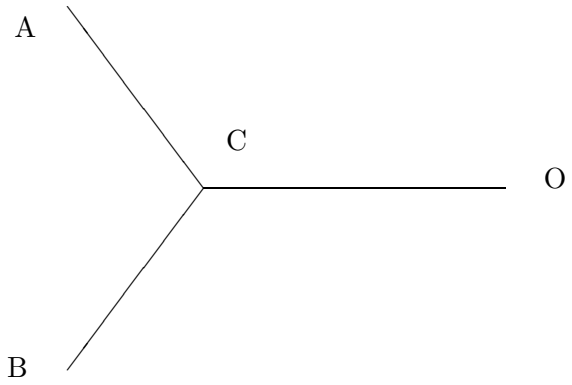


Figure 1.5: The unrooted phylogenetic tree of the species A, B, and O

that is, $H_0 : \mu_{AC}T_2 = \mu_{BC}T_2$. It is equivalent to testing $H_O : \mu_{AC} = \mu_{BC}$ by canceling the time T_2 out. One can first optimize the likelihood function of L without any constraint, say L_A, and then maximize L under the null hypothesis of equality of rates, say L_O. Then the likelihood ratio test can be constructed by

$$\lambda \quad = \quad L_O/L_A,$$

and $-2\log(\lambda)$ is approximate to be distributed by a χ^2-distribution with one degree of freedom. One of the benefits of LRT version of relative ratio test is that it can be easily applied to any substitution models.

1.3 Comparison of Patterns of Substitution Rates at Multiple Loci

1.3.1 Correlated Relative Rates among Loci

As the possibility of accessing multiple data sets of sequences for the same set of species became reality, various studies have been conducted to compare the relative rates of substitution at multiple genes or multiple genomes for the same set of species. This kind of study provides further insights into the mechanisms and processes of sequence evolution

and the functions of genes. For instance, if some evolutionary forces (*e.g.*, generation times or speciation rates) that affect the whole plant mitochondrial genome cause rate variation between lineages of a mitochondrial locus, we can expect to find the similar relative rates for most of mitochondrial loci. In other words, relative rates can be conserved among loci by a genome-wide evolutionary force. Conversely, if relative rates are not conserved well among loci, it indicates that locus-specific factors, say selection, predominate.

By comparing the plant mitochondrial, chloroplast, and nuclear sequences, Wolfe *et al.* (1987) showed that silent substitution rate in mtDNA was less 1/3 than that in cpDNA, which in turn evolved only 1/2 as fast as nDNA. Gaut and Clegg (1993)'s study gave the evidence that substitution rates in the maize *Adh1* lineage were 1.7 times faster than substitution rates in the millet *Adh1* lineage. The study by Muse and Gaut (1994) of the chloroplast genome suggested that relative rates were relatively well conserved across loci, and substitution rates were subject to both lineage-specific and locus-specific effects.

As previous studies suggested, two factors were considered to affect the patterns: lineage effects and locus effects. The lineage effects lead to changes in substitution rates at all loci in a particular evolutionary lineage. For example, organisms with 1/2 shorter generation times might lead to twice more substitution rates at all loci. Two distinct types of locus effects were taken into account. The first type is responsible for inter-locus rate differences that remain fairly constant across evolutionary lineages. For example, a previous study by Bulmer *et al.* (1991) indicated that observed variation in synonymous substitution rates among loci might result from regional differences in mutation rates. On the other hand, the second type of locus effect governs substitution rate at single locus in a particular lineage or set of lineages. Occasional intervals of locus-specific positive or negative selection can result in this type of effect. We might consider the latter one as an interaction between locus and lineage effects.

1.3.2 Relative Ratio Test

To answer the interesting question:"are relative substitution rates among particular lineages well conserved across genes?", relative ratio tests have been formulated. The relative ratio

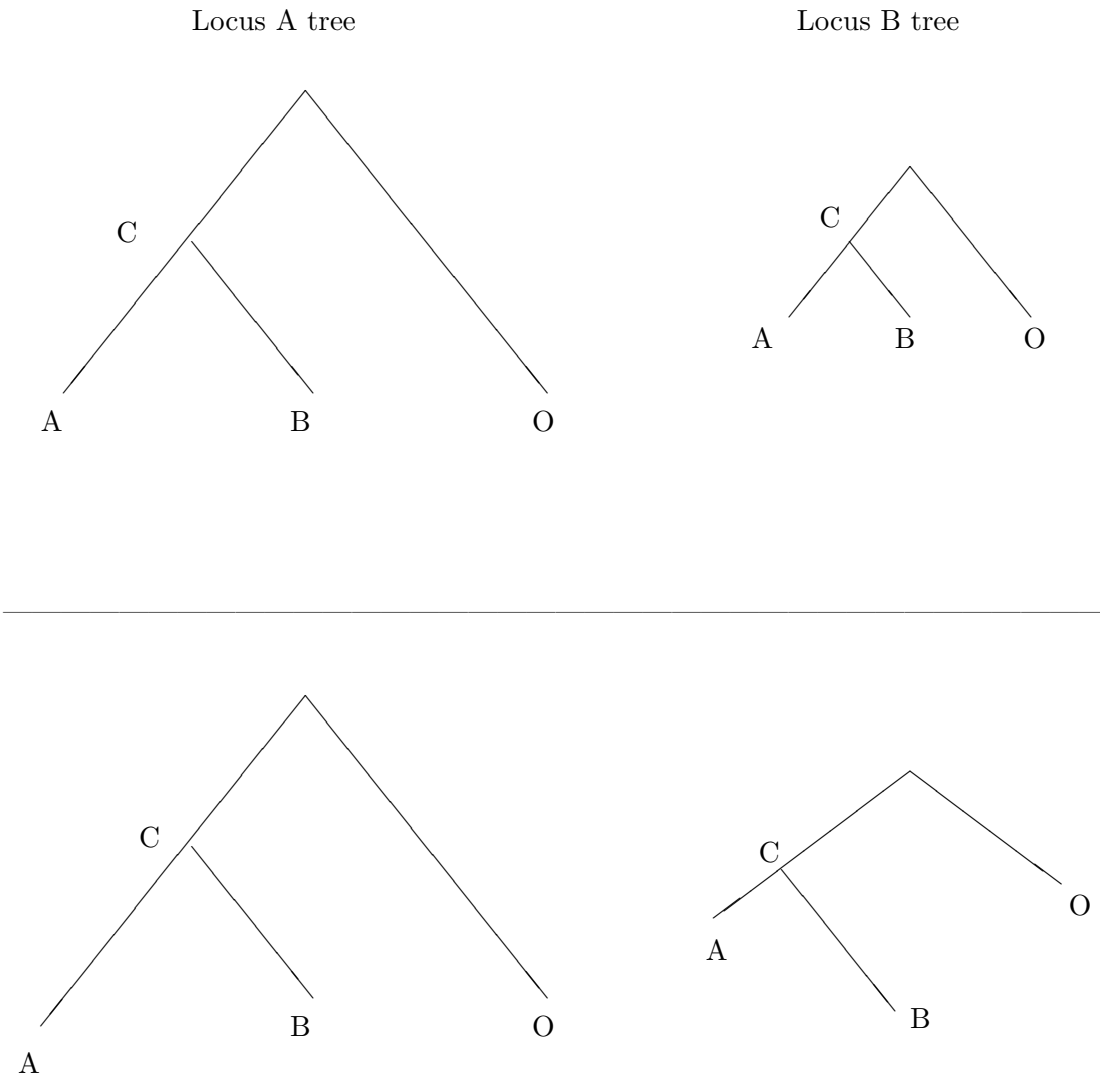


Figure 1.6: Comparing relative lengths of branches between two loci: branch lengths are generally longer in Locus A than in Locus B. In the top panel, although the absolute branch lengths are different between the two loci, the relative branch lengths are proportional. In the bottom panel, the two loci don't share the fast and slow branches.

test was designed for data from the same set of taxa for two or more genes. The purpose of the test is to determine whether or not the relative lengths of branches are conserved across loci in the same set of taxa. Consider the example in Figure 1.6. Branch lengths are generally longer in Locus A than in Locus B since Locus A evolves with a faster absolute rate than Locus B, but there are different scenarios between the top and bottom panels. In the top panel, although the absolute branch lengths are different between the two loci, the relative branch lengths are proportional. Such a proportion suggests that relative rates are conserved across loci, which can be explained by a single factor acting on both loci simultaneously. In the bottom panel, branch lengths are uncorrelated - the two loci don't share the fast and slow branches. As a consequence, no proportionality holds and relative rates are not conserved between loci. This requires different factors to be acted in the two loci.

Figure 1.7 illustrates testing hypotheses. Let the substitution rates for Locus A and Locus B on the branch i be denoted by μ_{A_i} and μ_{B_i}, respectively, and k be the proportionality constant. The hypotheses of interest are

$$\begin{aligned} H_0: &\quad \mu_{A_i} = k\mu_{B_i}, \\ H_a: &\quad \mu_{A_i} \neq k\mu_{B_i}. \end{aligned}$$

The relative rates are the same at the two loci if k is equal to 1. The constant k need not be known a priori, but can be estimated from the data. Based on the above discussions, the rejection of the null hypothesis indicates that the true substitution rates don't have proportional values at two loci, and that interplay of locus and lineage effects may be needed to explain differences.

Note that the hypotheses do not include time since the relative rates are independent of the time dimension, but time probably influences the performance of the test. Additionally, absolute rates can be different between loci without affecting the comparison. Different ways have been taken to implement the relative ratio test: distanced-based and maximum likelihood based. The former included tests developed by Eyre−Walker and Gaut (1997) and

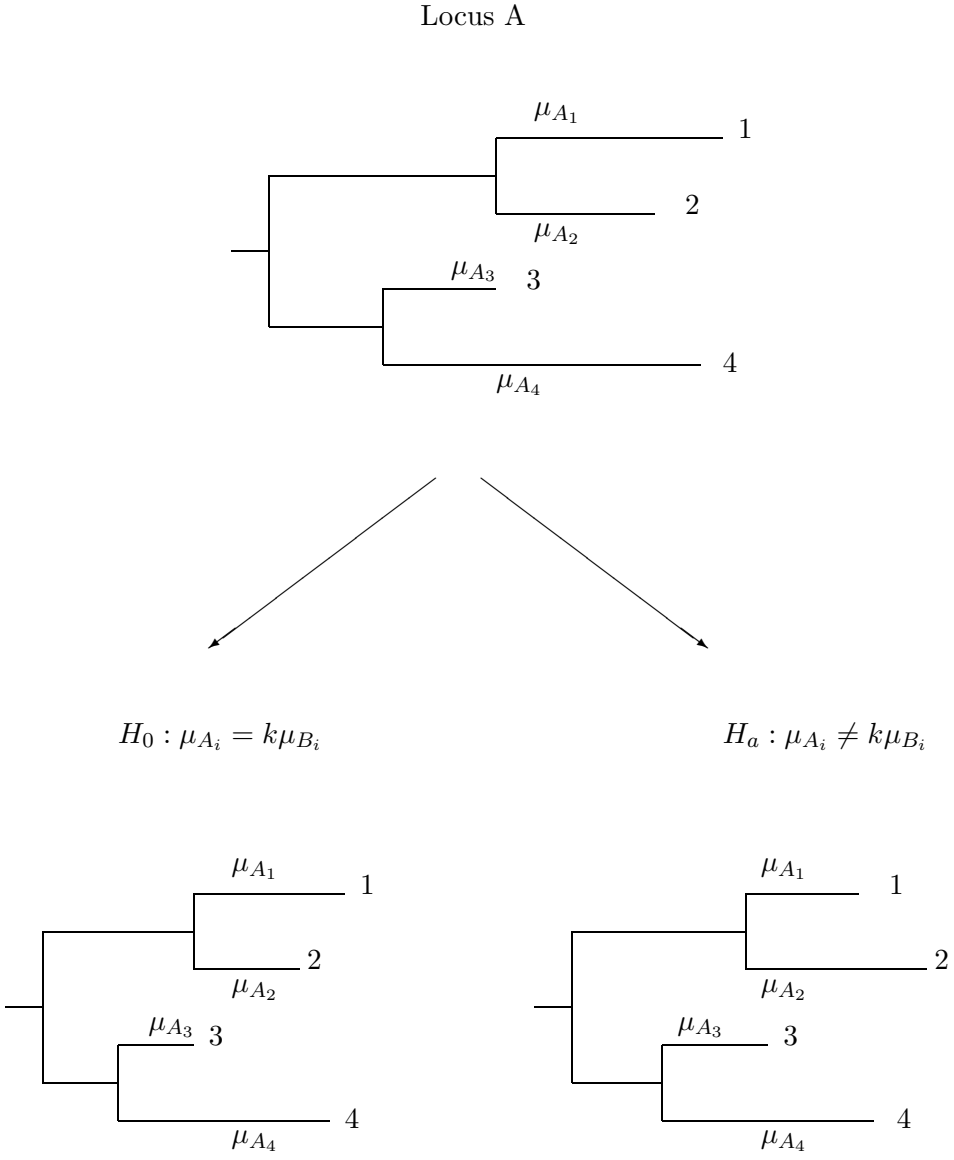


Figure 1.7: The Hypotheses of the relative ratio test: let the substitution rates for Locus A and Locus B on the branch i be denoted by μ_{A_i} and μ_{B_i}, respectively, and k be the proportionality constant.

Gaut *et al.* (1997); the latter was proposed by Muse and Gaut (1997). Brief introductions of these tests are given in the following subsections.

1.3.3 Distance-based Methods

The distance-based version of relative ratio test has been developed in two different ways by Eyre-Walker and Gaut (1997) and Gaut *et al.* (1997). In both procedures, the version of Li and Bousquet (1992)'s relative rate test has been used. As described previously, Li and Bousquet (1992)'s test is a distance-based method that compares substitution rates between groups of sequences by testing the average rates in two groups of sequences. Additionally, the test didn't make any assumption about relatedness of sequence groups, monophyletic or homogeneous.

Eyre-Walker and Gaut (1997) treated sequences in each locus as a group, and simply used Li and Bousquet (1992)'s relative rate test to detect whether average rates of substitution are equal between two loci or not. They estimated the branch lengths from a neighbor-joining tree (Saitou and Nei, 1987) and used bootstrapping to estimated the variances, then used Z-test to compare relative rates between loci.

Gaut *et al.*(1997) constructed a pairwise matrix of the ratio of substitution rates between two monophyletic groups of sequences for each locus and then used the Mantel test to detect whether there is any correlation between loci. The detailed procedure is described in the example of Figure 1.8. Suppose that two trees consist of two lineages each, with Lineage 1 having n taxa and Lineage 2 having m taxa. Let r be the ratio of substitution rates between two monophyletic groups of sequences contrasted in a relative rate test, which is defined as:

$$\hat{r} \quad = \quad \frac{\overline{d}_{10} + \overline{d}_{12} - \overline{d}_{20}}{\overline{d}_{20} + \overline{d}_{12} - \overline{d}_{10}},$$

where $\overline{d}_{10}$ is the average distance from n sequences in group one to the outgroup sequence, $\overline{d}_{20}$ is the average distance from m sequences in group two to the outgroup sequence, and $\overline{d}_{12}$ is the average distance of $n \times m$ distance pairs between two groups. Thus, pairwise matrix of $\hat{r}$ can be obtained for each locus. Mantel test (Sokal and Rolf 1995) was used to

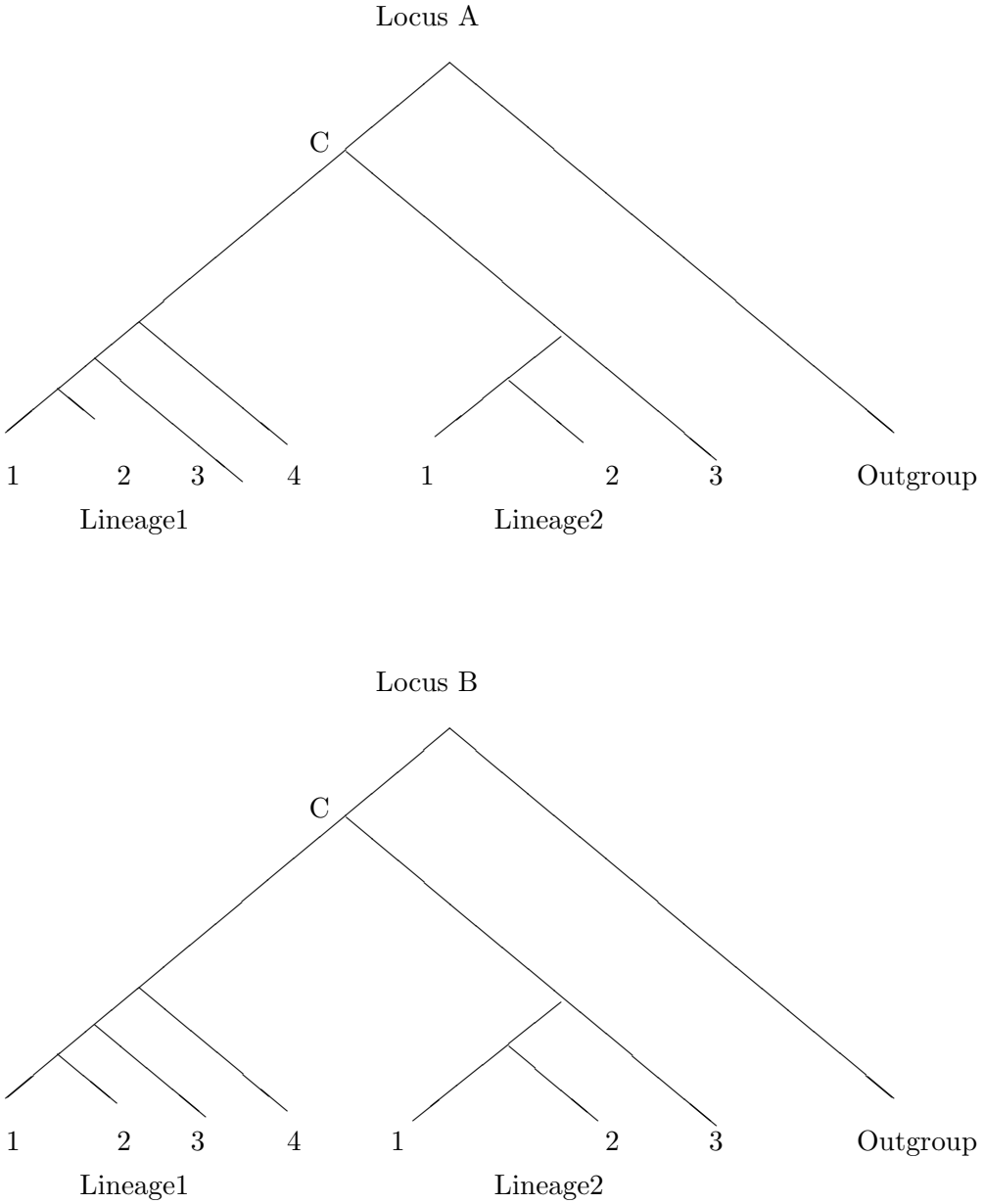


Figure 1.8: The phylogenetic tree of the distance-based relative ratio test of Gaut *et al.* (1997): two Locus, A and B, consist of three lineages each, with Lineage 1 having $n = 4$ taxa, Lineage 2 having $m = 3$ taxa and an outgroup

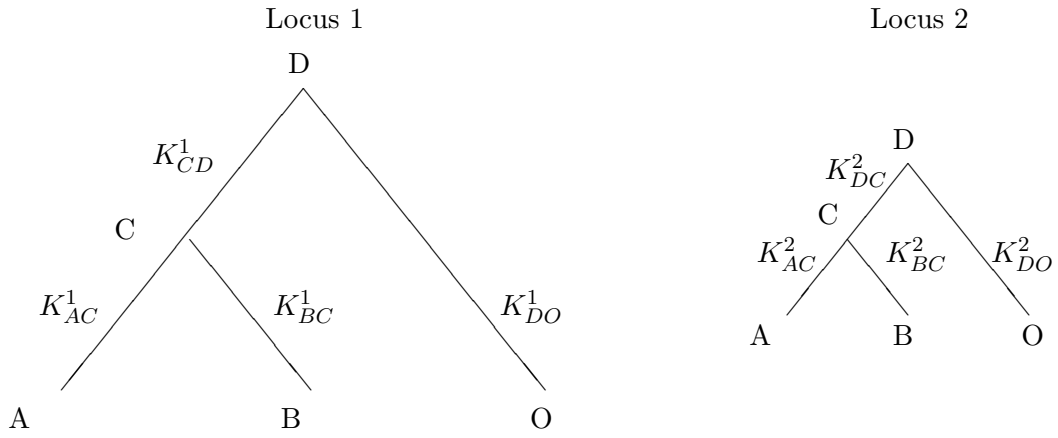


Figure 1.9: The Null hypothesis of the maximum likelihood-based relative ratio test: $K^1_{iC} = kK^2_{iC}$, where K^g_{ij} $(i, j = A, B, C, D, O,$ and $g = 1, 2)$ represents product of substitution rate and time for the species branch in the gth Locus.

test the correlation among the matrix of $\hat{r}$. A correlation coefficient, R_0, was calculated for the original matrices. They permutated the rows and columns within one of the matrices and recalculate the correlation coefficient. The measure of significance was the number of times that the original correlation coefficient (R_0) was exceeded by permutated values. They concluded that patterns of relative rates were correlated if the test is significant. A point made by Dr. Thorne that this procedure probably is not a real sense of relative ratio test since it doesn't explicitly test the null hypothesis of the relative ratio test.

1.3.4 Maximum Likelihood-based Method

The likelihood ratio version of the relative ratio test was proposed by Muse and Gaut (1997). Figure 1.9 illustrates the null hypothesis of the test. From the previous introduction, the likelihood function for site k_1, $k_1 = 1, \cdots, n$, at Locus 1 is given by

$$l^1_{k_1} = \sum_{S^1_{Ck_1}=1}^{4} \pi_{S^1_{Ck_1}} P_{S^1_{Ck_1},S^1_{Ok_1}}(K^1_{CO}) P_{S^1_{Ck_1},S^1_{Ak_1}}(K^1_{AC}) P_{S^1_{Ck_1},S^1_{Bk_1}}(K^1_{BC}),$$

and the likelihood function for site k_2, $k_2 = 1, \cdots, m$ at Locus 2 is given by

$$l^2_{k_2} = \sum_{S^2_{Ck_2}=1}^{4} \pi_{S^2_{Ck_2}} P_{S^2_{Ck_2},S^2_{Ok_2}}(K^2_{CO}) P_{S^2_{Ck_2},S^2_{Ak_2}}(K^2_{AC}) P_{S^2_{Ck_2},S^2_{Bk_2}}(K^2_{BC}),$$

where K^g_{ij} ($i, j = A, B, C, D, O$, and $g = 1, 2$) represents product of substitution rate and time for the species branch in the gth Locus. Let $L(K_1)$ and $L(K_2)$ denote the likelihood function for Locus 1 and Locus 2, respectively, we have

$$L(K_1) = \prod_{k_1=1}^{n} l^1_{k_1},$$

$$L(K_2) = \prod_{k_2=1}^{m} l^2_{k_2}.$$

Under the assumption that the substitution events are independent among the genes, the total likelihood function for both genes should be the product of two individual likelihood function. Without any constraint, the total likelihood function is given by

$$L_a = L(K_1)L(K_2)$$

Under the null hypothesis, the total likelihood function is given by

$$L_0 = L(kK_2)L(K_2)$$

So we can construct the likelihood ratio test as follows

$$\lambda(K) = \frac{L(kK_2)L(K_2)}{L(K_1)L(K_2)}$$

and $-2\log(\lambda(K))$ is approximately distributed as a χ^2-distribution with the degrees of freedom that depend on the number of compared branches.

Chapter 2

The Methods Based on Two-way ANOVA Model with Covariance Structure

2.1 Introduction

The analysis of variance is one of the most widely used statistical techniques. The basic idea of ANOVA is to partition variation in means. As mentioned in the previous chapter, the factors which govern variations in nucleotide substitutions can be considered as lineage effects, locus effects, and interaction between both. It is natural to use ANOVA method to deal with our problem. Applying ANOVA for testing the proportionality of relative rates at two loci is a method based on the use of pairwise sequence distances. Previous distance-based methods are valid, but they are probably not a powerful approach because they don't account for variations among branch lengths. From previous discussions, estimates of branch lengths within a locus are negatively correlated with each other since they are obtained by subtraction from estimates of the total number of substitutions separating two species. Therefore, the classic ANOVA won't be applicable for analyzing substitution rates, and the covariances among branch lengths have to be taken into account. It is not straightforward to obtain an analytical result for the covariance due to the complexity of substitution models. Fortunately, Bulmer (1989) derived the covariance of the number of substitutions between each pair of species for the model of Jukes and Canton (1969). The covariance between branch lengths can be easily deduced from it. Because of the existence of covariance between branch lengths, standard F tests are invalid. The expected mean squares

should be derived with incorporation of covariance. In addition to the aforementioned problems, several other problems are encountered when we perform the ANOVA and will be described in the following sections.

2.1.1 The Hypotheses of the ANOVA-based Tests

Figure 2.1 illustrates the data structure of interest. The question is whether two species, 1 and 2, have evolved at the same proportional rates at two loci, A and B, since they diverged from their common ancestor C. Note that the third species is an outgroup and used to root the tree. Write μ_{A_1}, μ_{A_2}, and μ_{A_3} for substitution rates for species 1, 2 and 3 at Locus A, respectively. Likewise, let μ_{B_1}, μ_{B_2}, and μ_{B_3} be substitution rates for species 1, 2, and 3 at Locus B. Our interest is to test whether or not the relative lengths of branches are proportional among loci, that is,

$$H_0 : \mu_{A_i} = k\mu_{B_i}, \qquad H_1 : \mu_{A_i} \neq k\mu_{B_i}.$$

Figure 2.2 illustrates four scenarios under no interaction between the lineage and locus effects. It is clear to see that testing the proportionality of relative rates is equivalent to testing no interaction from Figure 2.2. Thus the hypotheses of interest become

$$H_0 : \text{no interaction effect}, \qquad H_1 : \text{interaction effect exists}.$$

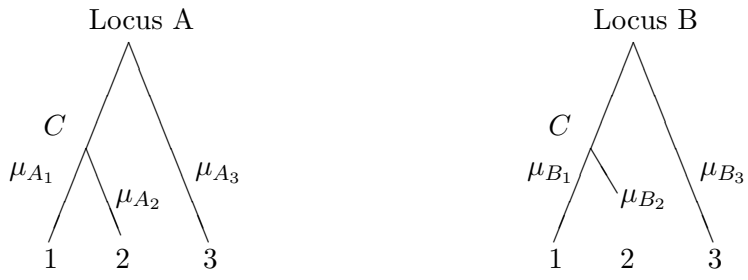


Figure 2.1: The tree structure of the ANOVA-based tests: whether two species, 1 and 2, have evolved at the same proportional rates at two loci, A and B, since they diverged from their common ancestor C. The third species is an outgroup and used to root the tree.

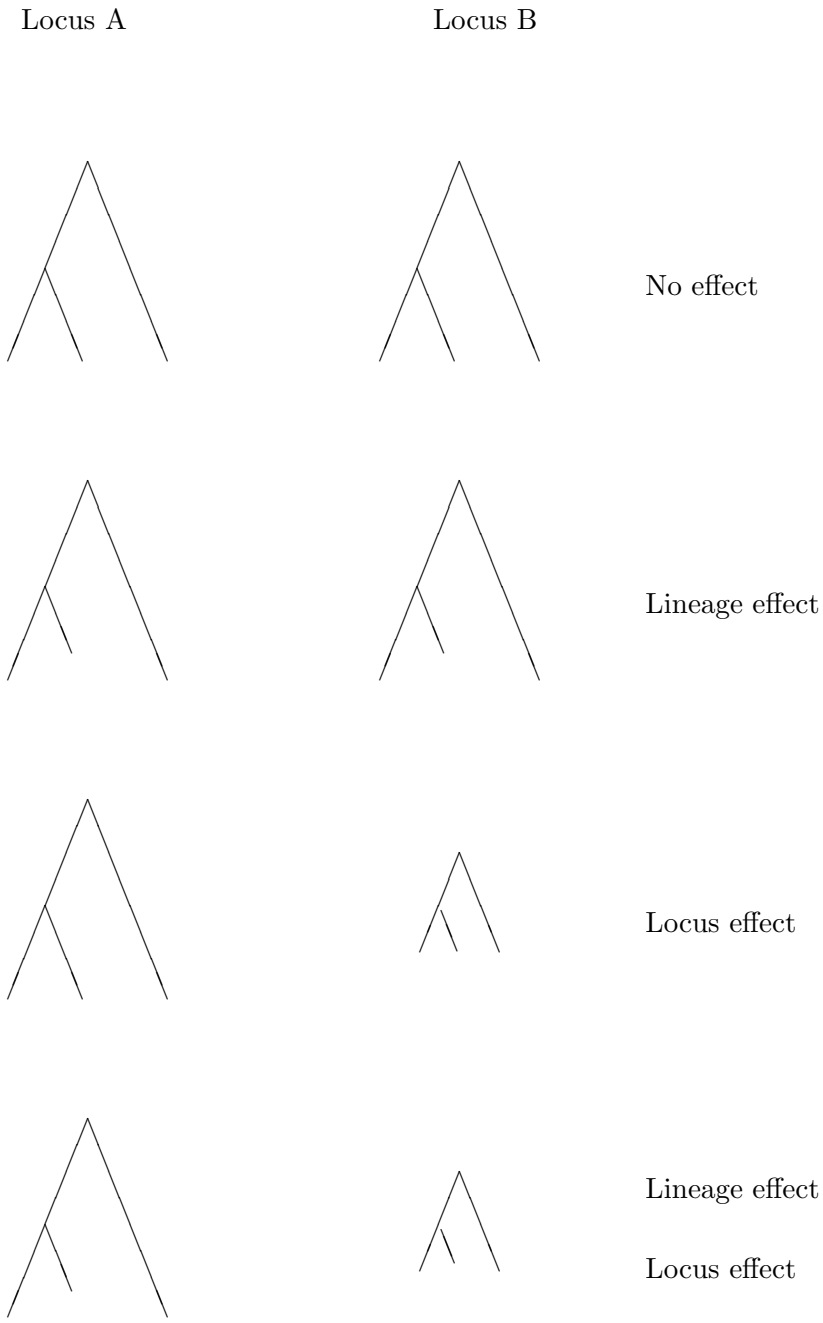


Figure 2.2: The four scenarios under no interaction between lineage and locus effects

2.1.2 Deriving Covariances between Branch Lengths

Bulmer (1989) evaluated variances and covariances of the expected number of substitutions between each pair of species by following the model of Jukes and Cantor (1969) in a formal argument. Suppose that s species from a homologous gene diverged from a common ancestor many years ago. Write μ_i for the expected number of substitutions per site in the ith lineage. Direct estimates of the μ_i's can not be made, but the number of substitutions between each pair of species, say d_{ij}, can be estimated. For species i and j, let p_{ij} represent the proportion of identical sites in the two homologous sequences. We have

$$\widehat{d}_{ij} \quad = \tfrac{3}{4} \ln \left(\tfrac{3}{4p_{ij}-1} \right).$$

Note that

$$\widehat{d}_{ij} \quad = \quad \mu_i + \mu_j + \varepsilon_{ij}.$$

Bulmer (1989) obtained the following formula:

$$\mathrm{Var}(\widehat{d}_{ij}) \quad = \tfrac{3}{4n} \left\{ \tfrac{1}{4} \exp \left[\tfrac{8}{3}(\mu_i + \mu_j) \right] + \tfrac{1}{2} \exp \left[\tfrac{4}{3}(\mu_i + \mu_j) \right] - \tfrac{3}{4} \right\}, \tag{2.1}$$

and

$$\mathrm{Cov}(\widehat{d}_{ij}, \widehat{d}_{ik}) \quad = \tfrac{3}{4n} \left[\tfrac{1}{4} \exp \left(\tfrac{8}{3}\mu_i \right) + \tfrac{1}{2} \exp \left(\tfrac{4}{3}\mu_i \right) - \tfrac{3}{4} \right], \tag{2.2}$$

where n is defined as the number of nucleotides in two homologous sequences being compared. Now consider three species. If the third is an outgroup, we will have

$$\hat{\mu}_1 \quad = \tfrac{1}{2}(\widehat{d}_{12} + \widehat{d}_{13} - \widehat{d}_{23}),$$

$$\hat{\mu}_2 \quad = \tfrac{1}{2}(\widehat{d}_{12} + \widehat{d}_{23} - \widehat{d}_{13}),$$

and

$$\begin{aligned}
\mathrm{Cov}(\hat{\mu}_1, \hat{\mu}_2) &= \frac{1}{4}\mathrm{Cov}\left[(\widehat{d}_{12} + \widehat{d}_{13} - \widehat{d}_{23}), (\widehat{d}_{12} + \widehat{d}_{23} - \widehat{d}_{13})\right] \\
&= \frac{1}{4}\Big[\mathrm{Cov}(\widehat{d}_{12}, \widehat{d}_{12}) + \mathrm{Cov}(\widehat{d}_{12}, \widehat{d}_{23}) - \mathrm{Cov}(\widehat{d}_{12}, \widehat{d}_{13}) \\
&\quad + \mathrm{Cov}(\widehat{d}_{13}, \widehat{d}_{12}) + \mathrm{Cov}(\widehat{d}_{13}, \widehat{d}_{23}) - \mathrm{Cov}(\widehat{d}_{13}, \widehat{d}_{13}) \\
&\quad - \mathrm{Cov}(\widehat{d}_{23}, \widehat{d}_{12}) - \mathrm{Cov}(\widehat{d}_{23}, \widehat{d}_{23}) + \mathrm{Cov}(\widehat{d}_{23}, \widehat{d}_{13})\Big] \\
&= \frac{1}{4}[\mathrm{Var}(\widehat{d}_{12}) - \mathrm{Var}(\widehat{d}_{13}) - \mathrm{Var}(\widehat{d}_{23}) + 2\mathrm{Cov}(\widehat{d}_{13}, \widehat{d}_{23})].
\end{aligned}$$

Note that

$$\begin{aligned}
\frac{1}{4}\mathrm{Var}(\widehat{d}_{12}) &= \frac{3}{16n}\left\{\frac{1}{4}\exp\left[\frac{8}{3}(\mu_1 + \mu_2)\right] + \frac{1}{2}\exp\left[\frac{4}{3}(\mu_1 + \mu_2)\right] - \frac{3}{4}\right\} \\
&= \frac{3}{64n}\left\{\exp\left[\frac{8}{3}(\mu_1 + \mu_2)\right] + 2\exp\left[\frac{4}{3}(\mu_1 + \mu_2)\right] - 3\right\},
\end{aligned}$$

$$-\frac{1}{4}\mathrm{Var}(\widehat{d}_{13}) = \frac{3}{64n}\left\{-\exp\left[\frac{8}{3}(\mu_1 + \mu_3)\right] - 2\exp\left[\frac{4}{3}(\mu_1 + \mu_3)\right] + 3\right\},$$

$$-\frac{1}{4}\mathrm{Var}(\widehat{d}_{23}) = \frac{3}{64n}\left\{-\exp\left[\frac{8}{3}(\mu_2 + \mu_3)\right] - 2\exp\left[\frac{4}{3}(\mu_2 + \mu_3)\right] + 3\right\},$$

and

$$\frac{2}{4}\mathrm{Cov}(\widehat{d}_{13}, \widehat{d}_{23}) = \frac{3}{64n}\left\{2\exp\left[\frac{8}{3}(\mu_3)\right] + 4\exp\left[\frac{4}{3}(\mu_3)\right] - 6\right\}.$$

Therefore,

$$\begin{aligned}
\mathrm{Cov}(\hat{\mu}_1, \hat{\mu}_2) &= \frac{1}{4}[\mathrm{Var}(\widehat{d}_{12}) - \mathrm{Var}(\widehat{d}_{13}) - \mathrm{Var}(\widehat{d}_{23}) + 2\mathrm{Cov}(\widehat{d}_{13}, \widehat{d}_{23})] \\
&= \frac{3}{64n}\left\{\exp\left[\frac{8}{3}(\mu_1+\mu_2)\right] + 2\exp\left[\frac{4}{3}(\mu_1+\mu_2)\right] - 3\right. \\
&\quad - \exp\left[\frac{8}{3}(\mu_1+\mu_3)\right] - 2\exp\left[\frac{4}{3}(\mu_1+\mu_3)\right] + 3 \\
&\quad - \exp\left[\frac{8}{3}(\mu_2+\mu_3)\right] - 2\exp\left[\frac{4}{3}(\mu_2+\mu_3)\right] + 3 \\
&\quad \left. +2\exp\left(\frac{8}{3}\mu_3\right) + 4\exp\left(\frac{4}{3}\mu_3\right) - 6\right\} \\
&= \frac{3}{64n}\left\{\exp\left[\frac{8}{3}(\mu_1+\mu_2)\right] + 2\exp\left[\frac{4}{3}(\mu_1+\mu_2)\right]\right. \\
&\quad - \exp\left[\frac{8}{3}(\mu_1+\mu_3)\right] - 2\exp\left[\frac{4}{3}(\mu_1+\mu_3)\right] \\
&\quad - \exp\left[\frac{8}{3}(\mu_2+\mu_3)\right] - 2\exp\left[\frac{4}{3}(\mu_2+\mu_3)\right] \\
&\quad \left. +2\exp\left(\frac{8}{3}\mu_3\right) + 4\exp\left(\frac{4}{3}\mu_3\right) - 3\right\} \qquad (2.3)
\end{aligned}$$

2.2 The Models and Tests

2.2.1 The Two-way ANOVA Models

To compare the patterns of substitution rates between two independently evolving lineages at two loci, rather than relying on a dating of the divergence time of two species, we use a reference or an outgroup species. The purpose of the outgroup species is to "root" the tree. Let Y_{ij} be the branch length for the jth species at the ith locus. We consider the following two-way ANOVA model

$$Y_{ij} = \mu + \alpha_i + \beta_j + \gamma_{ij} + \varepsilon_{ij},$$

where μ denotes the grand mean, α_i denotes the ith locus effect, β_j denotes the jth lineage effect, γ_{ij} denotes the interaction between the jth lineage effect and the ith locus effect, and ε_{ij} is a random effect.

There are only four observations for two loci, so it's obvious that there are no remaining degrees of freedom for ε_{ij}. Increasing observations is the only way to overcome this lack of degree of freedom. Under the assumption of nucleotide substitutions occurring independently at each site, we propose to alternatively split the DNA sequence data to "increase" the sample size, as illustrated by Table 2.1. Suppose we have the following DNA sequence ACCTGGATCCGATGCC. We want to split the data into two data sets: odd and even data sets. The nucleotides at the odd positions are indexed by number 1, and the nucleotides at the even positions are indexed by number 2. The nucleotides indexed by number 1 will be assigned to the first replicate, the odd data set, and the nucleotides indexed by number 2 will be assigned to the second replicate, the even data set. The purpose of interweaving replicates rather than blocking piece by piece is to avoid local rate heterogeneity within sequences. After we split sequence data, branch lengths for each replicate are calculated respectively.

Table 2.1: Illustration of splitting sequence data

Sequence	A	C	C	T	G	G	A	T	C	C	G	A	T	G	C	C
position	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16
Index	1	2	1	2	1	2	1	2	1	2	1	2	1	2	1	2

Writing Y_{ijk} for the branch length for the jth lineage and the kth replicate at the ith locus, the desirable model is

$$Y_{ijk} \quad = \quad \mu + \alpha_i + \beta_j + \gamma_{ij} + \varepsilon_{ijk}$$

for $i = 1, \cdots, a$; $j = 1, \cdots, b$; and $k = 1, \cdots, n$, where α_i, β_j and γ_{ij} represent locus effect, lineage effect and interaction between them, respectively. Assume that ε_{ijk} is distributed as a normal distribution with zero mean and proposed variances in the following subsection.

2.2.2 The Expected Mean Squares and F-tests

To obtain analytical results we first assume equality of variances for branch lengths at all lineages. This variance will be denoted by σ^2. Observed variations in substitution rates imply this assumption may not be realistic during evolutionary history. Research can get started from the simplest case. Moreover, the assumption of common variance is one of the important assumptions in the ANOVA. Secondly, we assume that DNA sequences along the same evolutionary lineage diverge independently across loci. That is, the branch lengths along the same lineage have zero covariances across loci. Thirdly, we assume that branch lengths are correlated to each other in the same magnitude within a locus. This covariance at the ith locus will be denoted by σ_{i12}. Relaxing these rather restrictive assumptions will be done in next chapters. As pointed out by Thorne *et al.* (1998), biologically speak, the substitution rates among closely related evolutionary lineages are probably correlated in a positive manner because those lineages are more likely to share the similar evolutionary forces during divergences. On the other hand, estimates of branch lengths within loci might be negatively correlated with each other since they are obtained by subtraction from estimates of the total number of substitutions per site separating two species. Therefore we assume that $\sigma_{i12} < 0$.

After we split the sequence data to "increase" observations, those assumptions are valid within each replicate. We also assume that there is no correlation among branch lengths across replicates. Let $\varepsilon_{i_1,j_1,k_1}$ be the random effect for the j_1th lineage and the k_1th replicate at the i_1th locus. Similarly, let $\varepsilon_{i_2,j_2,k_2}$ be the random effect for the j_2th lineage and the k_2th replicate at the i_2th locus. Let σ^2 and σ_{i12} be the same as the above. The summary

of the assumptions is illustrated as follows:

$$\begin{aligned}
&\mathrm{Cov}(\varepsilon_{i_1,j_1,k_1}, \varepsilon_{i_2,j_2,k_2}) \\
&= \begin{cases}
0, & \text{if } i_1 \neq i_2; \\
0, & \text{if } k_1 \neq k_2; \\
\sigma^2, & \text{if } i_1 = i_2,\ j_1 = j_2, \text{ and } k_1 = k_2; \\
\sigma_{i12} < 0, & \text{if } i_1 = i_2,\ j_1 \neq j_2, \text{ and } k_1 = k_2.
\end{cases}
\end{aligned}$$

Assume $E[\varepsilon_{ijk}] = 0$. Use $E[MSA]$, $E[MSB]$ and $E[MSAB]$ to denote expectation of locus mean square, lineage mean square and interaction mean square, respectively. Under the following restrictions:

$$\sum_{i=1}^{a} \alpha_i = 0, \quad \sum_{j=1}^{b} \beta_j = 0, \quad \sum_{i=1}^{a} \gamma_{ij} = 0, \quad \text{and} \quad \sum_{j=1}^{b} \gamma_{ij} = 0,$$

we then derive $E[MSA]$, $E[MSB]$ and $E[MSAB]$ (Appendix A)

$$E[MSA] = \frac{nb}{a-1} \sum_{i=1}^{a} \alpha_i^2 + \sigma^2 + \frac{b-1}{a} \sum_{i=1}^{a} \sigma_{i12},$$

$$E[MSB] = \frac{na}{b-1} \sum_{j=1}^{b} \beta_j^2 + \sigma^2 - \frac{1}{a} \sum_{i=1}^{a} \sigma_{i12},$$

$$E[MSAB] = \frac{n}{(a-1)(b-1)} \sum_{i=1}^{a} \sum_{j=1}^{b} \gamma_{ij}^2 + \sigma^2 - \frac{1}{a} \sum_{i=1}^{a} \sigma_{i12},$$

and

$$E[MSE] = \sigma^2.$$

These expectations show that if there are no locus main effects (i.e. all $\alpha_i = 0$), MSA and

MSE have the same expectation term σ^2 except for the term $\frac{b-1}{a}\sum_{i=1}^{a}\sigma_{i12}$; similarly, if there are no lineage effects, MSB and MSE have the same expectation term σ^2 except for the term $-\frac{1}{a}\sum_{i=1}^{a}\sigma_{i12}$. Finally, if there are no interaction (i.e. if all $\gamma_{ij}=0$), MSAB and MSE have the same expectation term σ_{i12} except for the term $-\frac{1}{a}\sum_{i=1}^{a}\sigma_{i12}$. To construct the standard F^* test statistics based on ideas of the classic two-way ANOVA, all the extra terms comparing with $E(MSE)$ should be subtracted from the corresponding expectations. We then obtain the following F^* test statistics.

$$F_A^* = (MSA - \frac{b-1}{a}\sum_{i=1}^{a}\hat{\sigma}_{i12})/MSE,$$

$$F_B^* = (MSB + \frac{1}{a}\sum_{i=1}^{a}\hat{\sigma}_{i12})/MSE,$$

and

$$F_{AB}^* = (MSAB + \frac{1}{a}\sum_{i=1}^{a}\hat{\sigma}_{i12})/MSE,$$

where

$$\begin{aligned} MSA &= \frac{nb}{a-1}\sum_{i=1}^{a}\left(\overline{Y}_{i..}-\overline{Y}_{...}\right)^2, \\ MSB &= \frac{na}{b-1}\sum_{j=1}^{b}\left(\overline{Y}_{.j.}-\overline{Y}_{...}\right)^2, \\ MSAB &= \frac{n}{(a-1)(b-1)}\sum_{i=1}^{a}\sum_{j=1}^{b}\left(\overline{Y}_{ij.}-\overline{Y}_{i..}-\overline{Y}_{.j.}+\overline{Y}_{...}\right)^2, \\ MSE &= \frac{1}{ab(n-1)}\sum_{i=1}^{a}\sum_{j=1}^{b}\sum_{k=1}^{n}\left(Y_{ijk}-\overline{Y}_{ij.}\right)^2, \end{aligned}$$

and the estimates of σ_{i12} can be obtained using (2.3). To test whether there are no locus

effects, no lineage effects and no interactions, we use the critical values from the F distribution. When null hypotheses hold, the decision rules to control the Type I errors at the level α are:

If $F_A^* \leq F_{(k-1)ab}^{(a-1)}(1-\alpha)$, does not reject H_0; otherwise, reject H_0,
If $F_B^* \leq F_{(k-1)ab}^{(b-1)}(1-\alpha)$, does not reject H_0; otherwise, reject H_0,
If $F_{AB}^* \leq F_{(k-1)ab}^{(a-1)(b-1)}(1-\alpha)$, does not reject H_0; otherwise, reject H_0.

2.3 The Type I Errors of the Tests

Generally, hypothesis tests are evaluated and compared through their probabilities of Type I error and power. In order to evaluate the proposed ANOVA-based relative ratio tests, simulations are performed to compare Type I errors, and powers between the LRT-based relative ratio tests and ANOVA-based relative ratio tests. The results of Type I errors are presented in this section, while the results of the powers will be given in next section.

Although we hope that the test statistics F^* described in the above will be approximately distributed as F distribution with appropriate degrees of freedom under the null hypotheses, we wish to check this for different sequence lengths in order to establish a set of working rules. To find the sequence length needed in order to invoke the asymptotic F distribution of the test statistics, two sets of 1000 replicate sets of three sequences with length $500bp$, $1000bp$ and $10000bp$ each, respectively, are generated by creating two ancestral sequences by different base frequencies and then allowing the sequences diverge independently by mutation according to the models of JC69 and HKY85, respectively. The simulation is conducted using the C++ programming language.

2.3.1 The Parameter Sets

A total of 6 DNA sequences is needed to be generated for three species at two loci for each simulation. The parameters used to generate those data include the base frequencies for the ancestor sequences, transition rates, and transversion rates for each species at two loci. In the simulation study of Type I errors, we considered the four sampling scenarios from

Figure 2.2. We let each parameter set represent different sampling schemes under the null hypothesis H_0.

set 1: no interaction, no locus effects and no lineage effects

set 2: no interaction, having locus effects and no lineage effects

set 3: no interaction, no locus effects and having lineage effects

set 4: no interaction, having locus effects and lineage effects.

We also consider a variety of combinations of base frequencies with transition and transversion rates. The cases of equal or unequal base frequencies are taken account under the equal or unequal transitional and transversional rates. We surveyed mitochondrial sequences in Genbank to obtain the proportion for the four bases in the cases of unequal base frequencies. In the cases of unequal rates, we let the transitional rates be twice of the transversional rates. There are four groups of parameter sets, each with four sampling schemes under the null hypothesis. The parameter sets will be tabulated as follows.

Table 2.2: The first group of parameter sets for ANOVA-based methods

Locus	Set	Base freq				Transition			Transversion		
		A	C	G	T	δ_1	δ_2	δ_3	ρ_1	ρ_2	ρ_3
A	1	.25	.25	.25	.25	.10	.10	.20	.10	.10	.20
	2	.25	.25	.25	.25	.10	.10	.25	.10	.10	.25
	3	.25	.25	.25	.25	.10	.20	.50	.10	.20	.50
	4	.25	.25	.25	.25	.05	.10	.25	.05	.10	.25
B	1	.25	.25	.25	.25	.10	.10	.20	.10	.10	.20
	2	.25	.25	.25	.25	.20	.20	.50	.20	.20	.50
	3	.25	.25	.25	.25	.10	.20	.50	.10	.20	.50
	4	.25	.25	.25	.25	.10	.20	.50	.10	.20	.50

Table 2.3: The second group of parameter sets for ANOVA-based methods

Locus	Set	Base freq				Transition			Transversion		
		A	C	G	T	δ_1	δ_2	δ_3	ρ_1	ρ_2	ρ_3
A	1	.25	.25	.25	.25	.20	.20	.40	.10	.10	.20
	2	.25	.25	.25	.25	.20	.20	.50	.10	.10	.25
	3	.25	.25	.25	.25	.10	.20	.50	.05	.10	.25
	4	.25	.25	.25	.25	.10	.20	.25	.05	.10	.125
B	1	.25	.25	.25	.25	.20	.20	.40	.10	.10	.20
	2	.25	.25	.25	.25	.40	.40	1.0	.20	.20	.50
	3	.25	.25	.25	.25	.10	.20	.50	.05	.10	.25
	4	.25	.25	.25	.25	.20	.40	.50	.10	.20	.25

Table 2.4: The third group of parameter sets for ANOVA-based methods

Locus	Set	Base freq				Transition			Transversion		
		A	C	G	T	δ_1	δ_2	δ_3	ρ_1	ρ_2	ρ_3
A	1	.40	.10	.20	.30	.05	.05	.10	.05	.05	.10
	2	.40	.10	.20	.30	.05	.05	.125	.05	.05	.125
	3	.40	.10	.20	.30	.05	.10	.25	.05	.10	.25
	4	.40	.10	.20	.30	.05	.10	.25	.05	.10	.25
B	1	.40	.10	.20	.30	.05	.05	.10	.05	.05	.10
	2	.40	.10	.20	.30	.10	.10	.25	.10	.10	.25
	3	.40	.10	.20	.30	.05	.10	.25	.05	.10	.25
	4	.40	.10	.20	.30	.10	.20	.50	.10	.20	.50

Table 2.5: The fourth group of parameter sets for ANOVA-based methods

Locus	Set	Base freq				Transition			Transversion		
		A	C	G	T	δ_1	δ_2	δ_3	ρ_1	ρ_2	ρ_3
A	1	.40	.10	.20	.30	.10	.10	.50	.05	.05	.25
	2	.40	.10	.20	.30	.10	.10	.25	.05	.05	.125
	3	.40	.10	.20	.30	.10	.20	.50	.05	.10	.25
	4	.40	.10	.20	.30	.10	.20	.25	.05	.10	.125
B	1	.40	.10	.20	.30	.10	.10	.50	.05	.05	.25
	2	.40	.10	.20	.30	.20	.20	.50	.10	.10	.25
	3	.40	.10	.20	.30	.10	.20	.50	.05	.10	.25
	4	.40	.10	.20	.30	.20	.40	.50	.10	.20	.25

2.3.2 The Primary Simulation

The primary simulation was conducted using sequences with total length 1000*bp* and 10000*bp*, respectively. Here we present the primary results for the sequences with 10000*bp*. After we generate simulated sequence data for species, first we alternatively split the sequence data for each species into 10, 20, 40 pieces as described before. Branch lengths are then calculated for each piece of sequence by using both the JC model and the HKY85 model. Relative ratio tests of the standard two-way ANOVA-based and the modified two-way ANOVA with incorporating covariances are performed for both the JC and HKY85 models. The likelihood version of relative ratio test is also conducted on the same data from both models in order to compare the performance among those procedures. The rejection rates of no interaction effects at 5% level are computed. The ANOVA represents the standard two-way ANOVA-based method, and ANOVA.C represents modified ANOVA with covariances. The results of Type I errors from the fourth group of parameter sets are tabulated in Table 2.6.

Table 2.6: Type I errors (%) of primary simulations for the sequences with length 10000*bp* under the 4th group

Model JC69							
Set	LRT	ANOVA			ANOVA.C		
		10	20	40	10	20	40
1	5.0	6.9	7.8	7.9	6.4	6.8	7.4
2	4.8	6.7	6.4	6.4	6.7	6.0	6.1
3	5.2	7.0	7.6	8.0	6.0	6.7	7.0
4	5.8	99.8	99.9	99.9	99.8	99.9	99.9
Model HKY85							
1	5.4	7.2	8.0	8.3	6.5	7.1	7.6
2	4.6	6.9	6.8	6.5	6.7	6.3	6.3
3	4.5	7.4	8.2	8.2	6.7	7.1	7.5
4	6.0	99.9	100.0	99.9	99.9	100.0	99.9

It is not surprising to see that rejection rates of the LRT-based method are much closer

to .05 than those of both ANOVA-based methods. The ANOVA.C improves the classic ANOVA for parameter set 1, set 2, and set 3, but the parameter set 4, having both locus and lineage effects, points out some unexpected results that all the rejection rates are almost 100% for both procedures. It is also interesting to note that consistent rejection rates within the same method no matter how big the replicates are. All simulation results are fairly consistent between the JC69 model and the HKY85 model.

2.3.3 The Simulation Results and Discussions

From the primary simulation study, it is clear that the new procedure doesn't fit all the cases, especially for the last sampling scheme that both locus effects and lineage effects exist. Many natural phenomena present on multiplicative scales, i.e., a system is more likely to "multiple" in response to a change of effects than to shift by a constant. As mentioned in Chapter 1, substitution rates tend to be inversely proportional to generation times. The following multiplicative model is probably more meaningful in the sense of biology,

$$Y_{ijk} \quad = \quad \alpha_i * \beta_j * \gamma_{ij} * \varepsilon_{ijk},$$

where Y_{ijk}, α_i, β_j, γ_{ij}, and ε_{ijk} are defined as before.

The transformation of logarithm is the natural method for analyzing data with additive model where the effects are thought to be multiplicative. It is clear that taking logarithms of branch lengths is the desirable way to transform the model to an ANOVA model. After taking natural log transformation, the model assumptions are the same as before except for the covariances between branch lengths. By the first Taylor expansion, covariances can be approximately computed by (See Appendix B)

$$\mathrm{Cov}(\ln \hat{\mu}_1, \ln \hat{\mu}_2) \quad \approx \quad \frac{1}{\mu_1 \mu_2} \mathrm{Cov}(\hat{\mu}_1, \hat{\mu}_2).$$

The total sequence lengths used to do simulations include $500bp$, $1000bp$, and $10000bp$. The $500bp$ and $1000bp$ sequences are split into 2, 5, and 10 replicates in an alternating way as described, respectively. We split the $10000bp$ sequence into 10, 20, and 40 replicates.

Let ANOVA be the standard procedure. The ANOVA.C represents the procedure of the standard two-way ANOVA with covariances. The ANOVA.L denotes the procedure of the standard two-way ANOVA after taking logarithms of branch lengths. Let the ANOVA.L.C stand for the procedure of the modified ANOVA after taking logarithms of branch lengths. Both JC69 and HKY85 are applied to the five tests - LRT, ANOVA, ANOVA.C, ANOVA.L, and ANOVA.L.C under the four groups of parameter sets. The Type I errors from the five procedures at the 0.05 level are tabulated together for each nucleotide model in order to compare the simulation results conveniently.

C++ programs have been written for simulations. The computational speeds are highly dependent on the total sequence lengths for both ANOVA-based and LRT-based procedures. The longer the total sequence is, the more time is needed. The speeds of ANOVA-based methods are 500 times faster than those of LRT-based procedures when HYPHY software is used to perform likelihood version of tests when programs are run under Sun Workstation. Another computational advantage for the ANOVA-based method is that it is easy to implement the hypothesis tests for both lineage and locus effects into the relative ratio test without increasing the computational time much. On the other hand, we should perform the likelihood version of tests three times to get the test results for locus effects, lineage effects, and the interaction between both. In this scenario, the ANOVA-based method will be 1500-fold faster than the likelihood method based on our computing environment.

Type I Errors Comparing with the ANOVA-based procedures, the LRT-based methods produce the smallest Type I errors for most cases in our simulations. In those exceptional cases, the ANOVA.L.C procedures are superior to the LRT. Although most of Type I errors from the LRT are very close to the desired 0.05 level, there are still around 27% cases that Type I errors are equal to or bigger than .055. For examples, Table 2.9 shows that Type I error of the LRT is 7.2% at the HKY85 model, which is bigger than those of the ANOVA-based methods.

Simulation results show that incorporating covariances reduces the false positive rate. Taking log before doing ANOVA-based methods dramatically decreases false positive rates

for parameter set 4 for all the cases. The ANOVA.L.C method works quite well for many cases with Type I errors being close to the desired .05. In addition, the ANOVA.L.C procedures result in smallest false positive rates among the ANOVA-based methods. Those findings make the ANOVA.L.C tests more appealing since previous researches suggested that existence of both locus and lineage effects was common in real data. This leads to concluding that doing the data transformation before performing ANOVA-based tests is an essential step. Note that results from both nucleotide models are fairly consistent even when the data are from the HKY85 model. In other words, the simplest model of nucleotide - JC69 is applicable to general cases in our simulations. This is very useful because we have explicit formula for branch lengths and their covariances for the JC model among the published substitution models. Within the same procedure of the ANOVA-based methods, no significant decrements are found in false positive rates as the number of replicates is increased.

Choice of Outgroup The proposed ANOVA.L.C procedure is a distance-based method, i.e., the nucleotide branch lengths for each species are first calculated by introducing an outgroup species. It is possible to obtain non-positive values for the branch lengths, which will lead to invalid observations for the method of ANOVA.L.C since we can't take logarithms of non-positive values. Those percentages of non-positive have been presented within parentheses in all the tables and will reduce reliability and power to significantly detect true differences. In our simulation the ourgroups are always arranged to be the ones with largest substitution rates among three species. Based on simulation results, not only does the number of non-logarithms depend on the choice of an outgroup, but also does rely on the number of replicates and the total sequence length, i.e., the length of sequence for each replicate. For the locus with sequence length over $500bp$, we should keep the length of sequence for each replicate to be at least over $200bp$.

2.4 The Powers of the Tests

We performed simulation to investigate the power of ANOVA.L.C-based and LRT-based relative ratio tests under different phylogenetic trees. As mentioned in the first chapter, the motivation of the ANOVA-based methods is to avoid computational burden from LRT-version relative ratio test for analyzing large DNA sequence data at multiple genes. Thus power study is conducted on total length 10000*bp*. First two ancestral sequences are generated by different base frequencies, and then allowing the sequences diverge along six paths, each with different substitution rates, independently by mutation. As shown in Table 2.19, the parameters used to investigate powers allow transition rates to be twice of transversion rates and unequal base frequencies. For the four parameter sets, besides increasing the branch lengths for species 2 at gene A, we keep branch lengths for all the other species as constant.

First of all, note that all powers increase close to 1 as the branch lengths of the second species at gene B slightly increase for both ANOVA.L.C and LRT-based methods. Any larger differences are virtually certain to be detected. For the ANOVA.L.C method, results of powers are fairly consistent with that of Type I errors for both the models HKY85 and JC69. Look closely at Figure 2.3, it is clear to see that powers from the ANOVA.L.C have no significant changes as the number of replicates for both the models HKY85 and JC69 are increased. Comparing the top panel with bottom panel in Figure 2.3, powers of the JC69 are almost the same as that of the HKY85. As shown in Figure 2.4, both ANOVA.L.C and LRT-version relative ratio tests perform equally well when applying in the model JC69, but when both tests are applied into the model HKY85, powers from LRT-based test decrease 20%. This suggests that the method of ANOVA.L.C is robust in terms of choosing substitution models.

Table 2.7: Comparing Type I errors (%) of ANOVA-based methods with those of LRT-based method for the sequences with length 500*bp* at the 1st group

Set	LRT	ANOVA			ANOVA.C			ANOVA.L			ANOVA.L.C		
		2	5	10	2	5	10	2	5	10	2	5	10
Model JC69													
1	5.2	6.9	6.0	6.8	6.6	5.4	6.2	7.2	5.8	5.6	6.9	5.4(3.0)	5.1(65.0)
2	5.3	8.8	8.6	8.7	7.2	7.5	7.7	7.5	6.4	8.6	6.7	6.0(3.7)	8.2(58.6)
3	5.4	8.5	9.3	9.6	7.5	7.8	7.4	8.9	7.7	7.5	7.6	6.9(22.7)	5.5(87.3)
4	5.6	17.7	21.1	21.5	16.2	19.0	19.6	7.5	5.6	5.5	6.6	5.4(26.0)	5.5(92.0)
Model HKY85													
1	4.3	7.1	6.2	6.9	6.3	5.7	6.2	7.6	6.3	5.3	7.1	5.9(3.5)	5.0(68.0)
2	4.7	10.5	10.4	11.1	8.9	9.1	10.3	8.5	8.0	8.4	7.8	7.3(7.4)	8.4(75.0)
3	5.1	9.9	11.4	11.3	8.6	9.7	9.7	11.0	8.4	7.9	9.3	7.3(39.6)	7.9(96.2)
4	5.2	20.4	23.9	25.0	18.2	22.3	22.7	7.8	6.1	5.4	7.0	5.9(36.0)	5.4(96.0)

Table 2.8: Comparing Type I errors (%) of ANOVA-based methods with those of LRT-based method for the sequences with length 500*bp* at the 2nd group

Set	LRT	ANOVA			ANOVA.C			ANOVA.L			ANOVA.L.C		
		2	5	10	2	5	10	2	5	10	2	5	10
Model JC69													
1	5.1	5.8	6.1	6.2	5.3	5.8	5.4	5.9	5.4	6.3	5.4	5.2(1.3)	5.6(41.6)
2	5.9	9.4	10.2	8.5	8.6	8.9	7.4	7.7	8.6	6.7	6.9	7.6(2.4)	6.4(48.6)
3	5.7	7.3	8.0	7.6	6.5	7.2	7.0	7.6	6.8	7.0	6.6	6.2(2.7)	6.6(91.6)
4	4.6	18.7	23.3	26.0	18.2	22.1	24.1	6.8	5.8	6.8	6.4	5.6(8.8)	6.1(71.5)
Model HKY85													
1	4.1	5.7	5.5	6.5	5.6	5.2	5.9	6.2	5.1	6.5	5.6	4.7(1.5)	6.3(46.0)
2	5.7	10.5	9.6	11.2	9.2	8.4	9.9	8.0	8.0	8.3	7.1	7.2(7.7)	8.3(80.8)
3	5.4	7.4	8.1	7.2	6.6	7.3	6.8	6.5	7.1	6.6	6.0	6.3(3.1)	6.2(93.5)
4	5.4	19.2	24.9	28.0	18.7	23.7	26.3	6.9	5.8	6.8	6.5	5.3(9.1)	6.2 (76.4)

Note: the numbers inside parentheses are percentages of non-positive branch lengths for the both methods of ANOVA.L and ANOVA.L.C.

Table 2.9: Comparing Type I errors (%) of ANOVA-based methods with those of LRT-based method for the sequences with length 500*bp* at the 3rd group

Model JC69													
Set	LRT	ANOVA			ANOVA.C			ANOVA.L			ANOVA.L.C		
		2	5	10	2	5	10	2	5	10	2	5	10
1	5.1	5.2	5.0	5.6	5.0	4.9	5.5	4.5	4.7	0.0	4.4	4.5(47.1)	0.0(99.9)
2	6.0	6.8	7.8	7.1	6.4	7.2	7.0	5.3	6.5	0.0	5.1	6.0(28.0)	0.0(98.6)
3	4.8	6.0	6.1	7.1	5.7	5.6	6.1	6.1	5.9	0.0	5.5	5.5(32.2)	0.0(98.0)
4	5.0	15.9	20.0	20.5	14.4	17.9	19.6	7.9	5.3	3.7	7.1	4.6(28.9)	1.9(94.6)
Model HKY85													
1	7.2	5.1	5.1	6.3	5.0	5.0	5.8	4.5	4.8	0.0	4.5	4.4(47.4)	0.0(99.9)
2	6.3	6.8	7.9	7.6	6.4	7.5	7.4	5.5	7.2	0.0	5.2	6.5(29.1)	0.0 (98.7)
3	5.4	6.3	6.6	7.5	5.8	6.1	6.9	6.2	6.1	0.0	5.7	5.8 (37.6)	0.0(98.7)
4	6.1	18.1	22.3	23.2	16.2	20.6	23.2	7.4	4.4	1.21	6.9	3.3 (36.9)	1.2(98.7)

Table 2.10: Comparing Type I errors (%) of ANOVA-based methods with those of LRT-based method for the sequences with length 500*bp* at the 4th group

Model JC69													
Set	LRT	ANOVA			ANOVA.C			ANOVA.L			ANOVA.L.C		
		2	5	10	2	5	10	2	5	10	2	5	10
1	5.3	7.7	6.7	7.1	6.7	6.1	6.2	6.8	4.5	0.0	6.1	4.4(28.7)	0.0(98.5)
2	5.6	6.7	7.3	6.9	6.4	6.6	6.7	6.0	5.7	8.9	5.8	5.6(10.6)	8.9(91.0)
3	4.1	6.3	6.0	6.8	5.6	5.1	6.0	6.7	5.4	4.6	5.8	4.6(28.1)	3.4(91.2)
4	5.9	17.5	22.3	23.5	16.8	21.5	22.6	6.6	6.6	9.0	6.2	5.7(8.9)	8.2(76.7)
Model HKY85													
1	5.8	7.8	6.8	7.8	6.7	6.3	6.7	6.7	4.9	0	6.1	4.6(33.2)	0.0(99.2)
2	4.6	7.0	7.3	7.3	6.5	7.0	6.6	5.9	5.9	5.5	5.8	5.9(12.2)	5.5(99.2)
3	4.9	6.2	6.2	7.2	5.8	5.6	6.6	6.7	5.3	5.0	5.6	4.9(32.5)	3.3(94.0)
4	7.0	18.4	24.0	25.9	17.7	23.0	24.6	6.7	6.3	8.1	6.1	6.1(11.3)	8.1(82.7)

Table 2.11: Comparing Type I errors (%) of ANOVA-based methods with those of LRT-based method for the sequences with length 1000*bp* at the 1st group

Set	LRT	ANOVA			ANOVA.C			ANOVA.L			ANOVA.L.C		
		2	5	10	2	5	10	2	5	10	2	5	10
Model JC69													
1	5.0	5.0	6.0	6.0	4.7	5.8	5.5	4.8	6.0	6.0	4.4	5.6	5.4(5.1)
2	4.8	8.1	7.8	7.2	6.8	6.9	6.5	7.4	7.0	6.2	6.5	6.4	5.5(10.5)
3	5.2	8.1	7.9	7.7	6.6	7.1	6.4	7.4	7.9	6.6	6.3	6.8	5.7(40.0)
4	4.4	25.4	34.4	36.6	23.6	32.0	33.1	5.9	6.2	5.4	5.2	5.2	4.5(47.0)
Model HKY85													
1	5.1	4.8	5.9	5.9	4.6	5.6	5.5	4.5	5.7	5.6	4.2	5.6	5.3(5.6)
2	4.2	9.3	8.5	9.5	8.4	8.0	8.8	8.3	7.6	7.6	7.5	6.7	6.9(12.6)
3	5.8	8.8	8.6	9.1	7.2	7.6	8.2	8.2	8.2	10.1	7.2	7.1	8.3(61.5)
4	5.0	28.5	39.4	41.0	27.0	36.8	39.0	6.7	6.7	5.2	6.2	6.3	5.0(59.8)

Table 2.12: Comparing Type I errors (%) of ANOVA-based methods with those of LRT-based method for the sequences with length 1000*bp* at the 2nd group

Set	LRT	ANOVA			ANOVA.C			ANOVA.L			ANOVA.L.C		
		2	5	10	2	5	10	2	5	10	2	5	10
Model JC69													
1	4.2	6.0	5.5	6.0	5.4	5.1	5.3	6.2	5.4	5.8	5.7	5.0	5.4(3.0)
2	5.3	8.8	8.5	9.2	7.0	7.1	7.4	7.1	7.3	7.9	5.9	6.4	7.0 (5.1)
3	4.4	5.6	7.1	7.0	5.2	6.3	6.5	6.0	5.9	6.2	5.0	5.5	2.2(4.4)
4	5.0	30.2	41.6	43.8	29.4	39.8	42.8	5.6	5.4	4.6	5.0	5.1	4.6 (14.6)
Model HKY85													
1	4.6	5.9	6.2	6.0	5.2	5.6	5.6	6.2	6.3	6.0	5.7	5.9	5.5(4.5)
2	5.7	9.4	9.3	10.8	8.0	8.0	9.4	7.1	8.0	8.7	5.9	7.4	7.5(16.9)
3	4.7	5.6	7.2	7.1	5.2	6.4	6.7	6.5	6.0	5.4	5.0	5.1	5.2(4.8)
4	5.3	30.7	43.4	45.3	29.4	41.9	44.2	5.6	5.2	4.5	5.4	4.9	4.5(18.1)

Table 2.13: Comparing Type I errors (%) of ANOVA-based methods with those of LRT-based method for the sequences with length 1000*bp* at the 3rd group

Set	LRT	ANOVA			ANOVA.C			ANOVA.L			ANOVA.L.C		
		2	5	10	2	5	10	2	5	10	2	5	10
Model JC69													
1	6.2	6.1	5.9	6.7	5.7	5.7	6.1	5.4	5.8	6.8	5.2	5.8(1.6)	6.8(69.3)
2	5.4	6.3	6.0	5.9	6.1	5.2	5.6	5.3	5.7	5.8	5.1	5.5(1.3)	5.5(45.2)
3	4.7	6.6	6.0	6.6	6.0	5.5	5.9	7.2	6.2	6.0	6.7	5.8(4.5)	6.0(53.2)
4	6.3	22.9	32.8	32.8	21.3	30.4	30.8	8.1	7.8	7.8	7.3	7.0(2.9)	7.4(44.9)
Model HKY85													
1	5.6	6.1	6.2	6.7	6.0	6.0	6.4	5.3	5.9	6.9	5.2	5.6(1.6)	6.9(69.6)
2	4.6	6.5	5.8	6.3	6.2	5.5	6.1	5.3	5.5	6.0	5.0	5.5(1.5)	6.0(46.2)
3	4.7	6.4	6.6	6.7	6.1	5.9	6.4	7.5	6.3	4.8	7.0	6.1(6.3)	4.8(60.0)
4	6.0	25.9	36.1	37.3	25.0	34.3	35.4	8.2	7.9	7.6	7.3	7.6(5.9)	7.4(60.6)

Table 2.14: Comparing Type I errors (%) of ANOVA-based methods with those of LRT-based method for the sequences with length 1000*bp* at the 4th group

Set	LRT	ANOVA			ANOVA.C			ANOVA.L			ANOVA.L.C		
		2	5	10	2	5	10	2	5	10	2	5	10
Model JC69													
1	5.6	7.5	7.8	8.0	6.9	7.0	6.6	7.4	6.7	8.2	6.7	6.2(1.5)	8.2(48.7)
2	5.8	6.9	8.0	6.9	6.2	6.9	6.3	6.9	7.1	6.7	6.1	6.8(0.0)	6.7(21.2)
3	4.9	6.4	7.4	7.6	5.6	7.0	6.9	6.1	7.2	7.1	5.2	6.7(3.2)	6.2(42.3)
4	6.4	27.5	35.0	37.3	26.4	33.0	35.7	7.5	7.8	8.3	6.6	7.3(0.0)	8.2(17.8)
Model HKY85													
1	5.2	7.6	8.0	8.2	6.8	7.1	7.5	7.5	7.1	8.6	6.8	6.6(2.4)	7.9(54.4)
2	5.4	7.1	7.9	7.1	6.6	7.3	6.8	7.0	7.4	7.0	6.2	7.1(0.0)	6.9(22.9)
3	4.3	6.4	7.7	7.9	5.9	7.2	7.0	6.0	7.6	7.2	5.4	7.2(4.0)	6.9(50.9)
4	5.3	28.4	38.0	40.1	28.0	36.9	39.1	7.5	7.8	8.3	6.8	7.3(0.0)	8.0(22.7)

Table 2.15: Comparing Type I errors (%) of ANOVA-based methods with those of LRT-based method for the sequences with length 10000*bp* at the 1st group

Set	LRT	ANOVA			ANOVA.C			ANOVA.L			ANOVA.L.C		
		10	20	40	10	20	40	10	20	40	10	20	40
Model JC69													
1	4.7	6.1	5.8	5.4	5.6	5.5	5.2	5.9	6.3	5.4	5.5	5.4	5.0(0.0)
2	4.7	9.3	8.9	8.5	7.8	8.0	7.3	8.6	8.7	8.3	7.7	7.8	7.6(0.0)
3	5.2	8.1	8.0	7.7	6.7	6.5	6.4	8.1	7.8	7.5	7.3	6.9	6.7(16.3)
4	6.0	99.2	99.3	99.4	99.2	99.3	99.4	6.5	6.4	6.6	6.0	6.0	6.1(12.1)
Model HKY85													
1	5.3	6.0	5.6	5.7	5.4	5.3	5.1	6.1	5.9	6.6	5.7	5.4	6.0(0.0)
2	5.0	10.2	9.1	9.6	9.2	7.9	8.4	9.3	9.6	8.8	8.3	8.3	7.9 (0.0)
3	5.4	9.1	8.3	8.8	7.7	7.1	7.2	8.5	8.3	7.0	7.5	6.8	6.3(20.2)
4	5.7	100	100	99.8	100	100	99.8	8.8	8.5	9.0	8.1	8.1	8.6(17.0)

Table 2.16: Comparing Type I errors (%) of ANOVA-based methods with those of LRT-based method for the sequences with length 10000*bp* at the 2nd group

Set	LRT	ANOVA			ANOVA.C			ANOVA.L			ANOVA.L.C		
		10	20	40	10	20	40	10	20	40	10	20	40
Model JC69													
1	5.0	7.0	7.0	7.2	6.4	6.2	6.7	7.0	7.2	7.4	6.4	6.6	6.7(0.0)
2	4.8	9.2	8.9	8.3	7.7	7.6	7.2	8.5	8.3	8.1	7.4	7.1	6.7 (0.0)
3	4.3	6.2	6.9	6.6	5.5	6.2	5.8	6.5	7.0	6.3	6.0	6.2	5.6(6.8)
4	3.8	100	100	100	100	100	100	4.3	4.7	4.8	4.3	4.4	4.5(0.2)
Model HKY85													
1	4.4	6.5	6.2	7.0	5.8	5.7	6.2	6.4	6.4	7.0	5.9	5.9	6.6(0.0)
2	5.6	9.2	8.4	8.1	7.9	6.9	6.4	8.1	7.1	7.2	7.0	6.0	6.2(0.0)
3	4.2	6.5	6.7	6.7	5.8	6.2	6.0	6.3	7.0	6.1	5.6	6.0	5.7(8.5)
4	4.4	100.0	99.9	100.0	100.0	99.9	94.4	5.0	5.0	4.7	4.6	4.5	4.7(0.4)

Table 2.17: Comparing Type I errors (%) of ANOVA-based methods with those of LRT-based method for the sequences with length 10000*bp* at the 3rd group

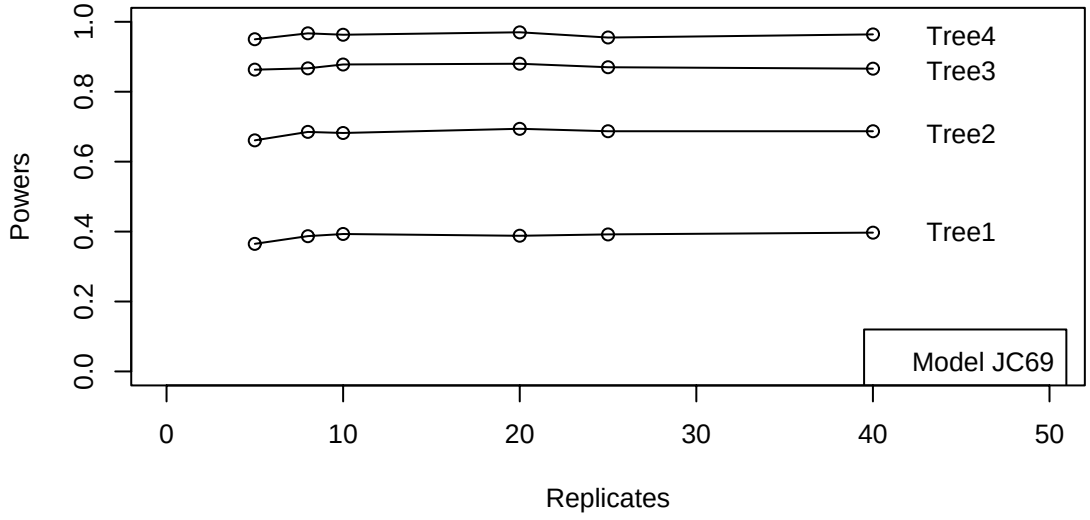
Set	LRT	ANOVA			ANOVA.C			ANOVA.L			ANOVA.L.C		
		10	20	40	10	20	40	10	20	40	10	20	40
Model JC69													
1	4.7	5.6	5.1	5.2	5.4	5.0	5.1	5.6	5.7	5.4	5.2	5.7	5.3(4.6)
2	5.4	7.4	7.8	7.3	7.3	7.6	7.1	7.3	6.5	6.2	7.0	6.4	5.9(1.4)
3	5.5	6.7	6.7	6.2	6.3	6.1	6.1	7.3	7.4	6.3	6.5	6.9	5.9(14.0)
4	4.9	98.2	98.7	98.8	97.8	98.5	98.3	7.4	8.0	8.4	6.9	6.9	7.4(9.2)
Model HKY85													
1	4.5	5.6	5.4	5.2	5.5	5.0	5.2	5.6	5.7	5.5	5.3	5.5	5.4(4.4)
2	4.8	7.7	7.9	8.2	7.3	7.5	7.1	7.3	7.0	6.7	7.1	6.5	6.0(1.6)
3	5.0	6.7	6.7	6.5	6.6	6.6	6.3	7.3	7.4	5.8	6.9	6.9	5.5(19.3)
4	5.8	99.4	99.5	99.6	99.4	99.4	99.6	7.7	7.5	7.4	7.0	6.9	6.9(17.9)

Table 2.18: Comparing Type I errors (%) of ANOVA-based methods with those of LRT-based method for the sequences with length 10000*bp* at the 4th group

Set	LRT	ANOVA			ANOVA.C			ANOVA.L			ANOVA.L.C		
		10	20	40	10	20	40	10	20	40	10	20	40
Model JC69													
1	5.0	6.9	7.8	7.9	6.4	6.8	7.4	6.8	8.0	8.0	6.3	7.4	7.6(3.9)
2	4.8	6.7	6.4	6.4	6.7	6.0	6.1	5.8	6.3	6.0	5.6	6.2	5.8(0.0)
3	5.2	7.0	7.6	8.0	6.0	6.7	7.0	6.4	6.8	6.8	5.9	6.4	6.1(6.1)
4	5.8	99.8	99.9	99.9	99.8	99.9	99.9	7.2	7.1	7.2	6.7	7.0	6.9(0.0)
Model HKY85													
1	5.4	7.2	8.0	8.3	6.5	7.1	7.6	6.8	8.1	8.5	6.3	7.6	8.0(5.3)
2	4.6	6.9	6.8	6.5	6.7	6.3	6.3	6.2	6.3	6.1	5.6	6.2	6.0(0.0)
3	4.5	7.4	8.2	8.2	6.7	7.1	7.5	6.8	6.9	6.8	6.3	6.7	5.9(8.0)
4	6.0	99.9	100.0	99.9	99.9	100.0	99.9	6.8	7.0	6.9	6.1	6.7	6.7(0.0)

Table 2.19: The parameter sets of power study for ANOVA-based methods

Locus	Tree	Base freq				Transition			Transversion		
		A	C	G	T	δ_1	δ_2	δ_3	ρ_1	ρ_2	ρ_3
A	1	.40	.10	.20	.30	.10	.10	.25	.05	.05	.15
	2	.40	.10	.20	.30	.10	.10	.25	.05	.05	.15
	3	.40	.10	.20	.30	.10	.10	.25	.05	.05	.15
	4	.40	.10	.20	.30	.10	.10	.25	.05	.05	.15
B	1	.40	.10	.20	.30	.10	.12	.25	.05	.060	.15
	2	.40	.10	.20	.30	.10	.13	.25	.05	.065	.15
	3	.40	.10	.20	.30	.10	.14	.25	.05	.070	.15
	4	.40	.10	.20	.30	.10	.15	.25	.20	.075	.15



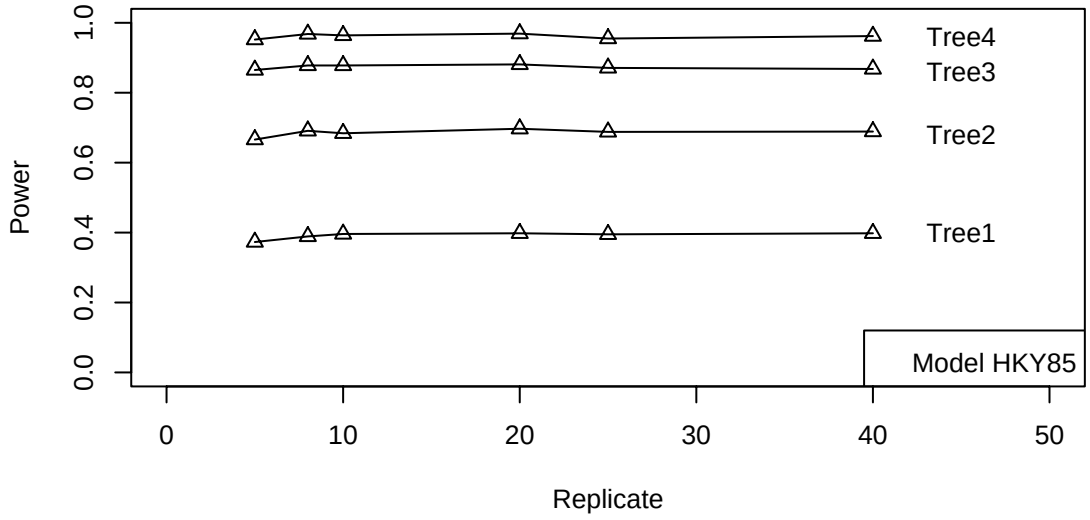
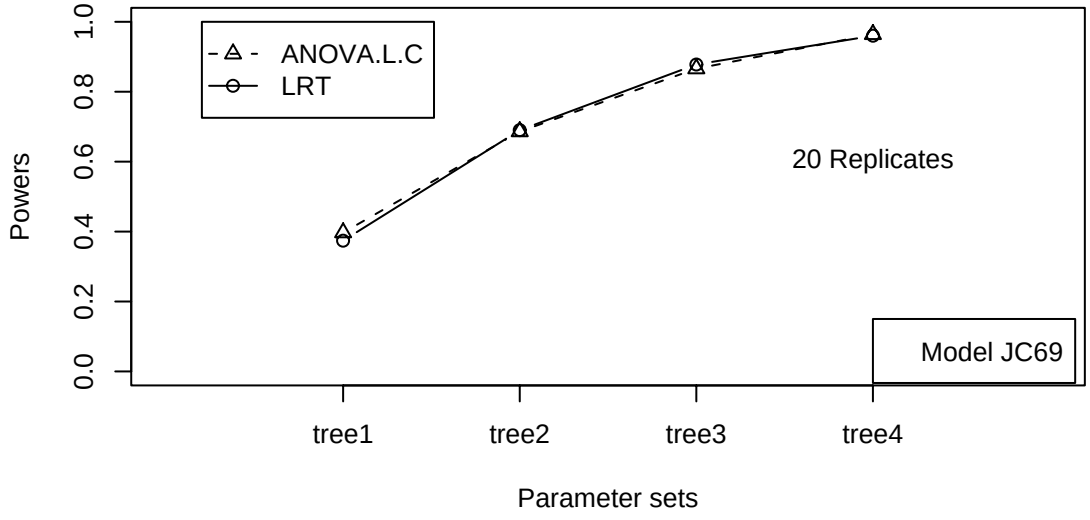


Figure 2.3: Power (%) study of ANOVA.L.C method



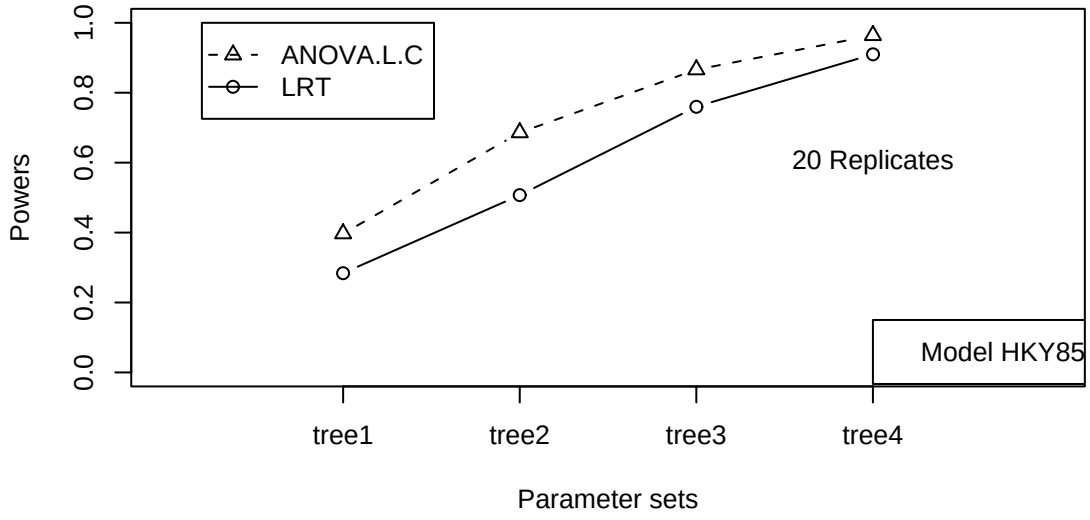


Figure 2.4: Comparing powers (%) between ANOVA.L.C and LRT

Chapter 3

The Methods Based on Generalized Estimating Equations

3.1 Introduction

The events of nucleotide substitution over sites are commonly assumed to follow Poisson process, which leads to using the Poisson distribution to describe the expected number of nucleotide substitutions per site. The simulation study in the previous chapter suggests the multiplicative model of substitution rates is more suitable to analyze sequence data as well as meaningful in biology. We propose to use Poisson regression in log-linear model, one type of generalized linear model, to analyze the substitution data. Since substitution rates are correlated, regular GLM which aims at uncorrelated data won't be appropriate. Liang and Zeger (1986) introduced generalized estimating equations (GEEs) as a method of dealing with correlated data when, except for the correlation among responses, the data can be modeled as a generalized linear model. One advantage of GEEs is to avoid making homogeneity of variance as well as the same covariance assumptions for substitution data analysis as in the ANOVA-based methods. To better understand the proposed method, a brief introduction of GLMs and GEEs will be given as follows.

3.1.1 Definition of GLM

Generalized linear models (GLM) are an extension of traditional linear models that allows the response probability distribution to be any member of an exponential family of distributions. It is flexible to use generalized linear models for a wider range of data analysis,

particularly for fitting counts and measured proportions data. In addition to the classical linear models with normal distributions, many other widely-used statistical models belong to generalized linear models, such as logistic and probit models for binomial variates, and log-linear models for multinomial and Poisson data.

In generalized linear models, response variables are assumed to be in a subclass of the one-parameter exponential family. That is, the probability density of a random variable $\boldsymbol{Y}$ for continuous responses, or the probability mass function for the discrete responses, can be written as

$$f(y|\theta;\phi) \quad = \quad \exp\left[\frac{\theta y - r(\theta)}{a(\phi)}\right] h(\phi, y), \tag{3.1}$$

for some functions r, a, and h which define various distributions. θ is an unknown parameter, and ϕ is an unknown constant. The function $a(\phi)$ is a positive function that allows ϕ to be defined easily for different cases. The expressions of mean and variance of response variable $\boldsymbol{Y}$ can be derived from standard theory. We have

$$\mathrm{E}(\boldsymbol{Y}) \quad = \quad r'(\theta) \equiv \mu, \tag{3.2}$$

$$\mathrm{Var}(\boldsymbol{Y}) \quad = \quad r''(\theta), \tag{3.3}$$

where the primes denote derivatives with respect to θ. Typically, r' is an invertible function so we can write θ as a function of the mean μ, say,

$$\theta = r_*(\mu).$$

For the ith response, a linear structure for distribution (3.1) is naturally written as

$$\theta_i = \boldsymbol{x}_i'\boldsymbol{\beta},$$

where $\boldsymbol{x}_i' = (x_{i1}, \cdots, x_{ip})$ is a $1 \times p$ row vector of known predictor variables, and $\boldsymbol{\beta} = (\beta_1, \cdots, \beta_p)'$ is a $p \times 1$ column vector of unknown parameters. Note that with $\theta_i = r_*(\mu_i)$,

the linear structure $\boldsymbol{x}_i'\boldsymbol{\beta}$ is also a function of the mean μ_i, more generally say,

$$g(\mu_i) = \boldsymbol{x}_i'\boldsymbol{\beta}. \tag{3.4}$$

Comparing expression (3.4) with (3.2) and (3.3), the variance of $\boldsymbol{Y}$ is also a function of the mean μ. It is often written as

$$v(\mu) = r''(\theta),$$

and $v(\mu)$ is often called as the variance function.

In summary a generalized linear model consists of independent responses y_i, $i = 1, \cdots, n$ with

$$y_i \sim f(y_i|\theta_i, \phi), \quad \mathrm{E}(y_i) \equiv \mu_i, \quad g(\mu_i) = \boldsymbol{x}_i'\boldsymbol{\beta}.$$

If $g(\mu_i) = \theta_i$, then the model is a canonical generalized linear model.

Conventionally, we call the linear structure $\boldsymbol{x}'\boldsymbol{\beta}$ as the linear predictor, while the function $g(\cdot)$ is the link function which specifies the relationship between the mean and linear predictor. If $g(\mu) = \theta$, the function $g(\cdot)$ is called the canonical link function. For instance, the identity is the canonical link for the normal distribution, the logit is the canonical link for the binomial distribution, and the log is the canonical link for the Poisson distribution. An advantage of canonical links is that a minimal sufficient statistic for $\boldsymbol{\beta}$ exists, i.e. all the information about $\boldsymbol{\beta}$ is contained in a function of the data of the same dimensionality as $\boldsymbol{\beta}$. Maximum likelihood functions are used to estimate parameters in generalized linear models.

3.1.2 Generalized Estimating Equations

A generalized linear model described above can be applied to uncorrelated data. Liang and Zeger (1986) introduced Generalized Estimating Equations (GEEs) to handle correlated

data when the data can be modeled as a generalized linear model. Instead of using a maximum likelihood, a class of estimating equations which take correlation into account were proposed to estimate the regression parameters and their variance-covariance matrices. The most common form of correlated data in the applied world is longitudinal data that include repeated measurements for each subject. The brief review of GEEs is illustrated using longitudinal data. Let Y_{ij} ($i = 1, \cdots, k$, and $j = 1, \cdots, n_i$) represent the jth measurement on the ith subject. Let the vector of measurements on the ith subject be $\boldsymbol{Y}_i = [Y_{i1}, \cdots, Y_{in_i}]'$ with corresponding vector of means $\boldsymbol{\mu}_i = [\mu_{i1}, ..., \mu_{in_i}]'$, and let $\boldsymbol{V}_i$ be an estimate of the covariance matrix of $\boldsymbol{Y}_i$. Let the vector of independent variables for the jth measurement on the ith subject be

$$\boldsymbol{x}_{ij} \quad = \quad [x_{ij1}, ..., x_{ijp}]'$$

The Generalized Estimating Equation for estimating $\boldsymbol{\beta}$ is given by

$$\sum_{i=1}^{K} \frac{\partial \boldsymbol{\mu}_i'}{\partial \boldsymbol{\beta}} \boldsymbol{V}_i^{-1} \left(\boldsymbol{Y}_i - \boldsymbol{\mu}_i(\boldsymbol{\beta}) \right) \quad = \quad \mathbf{0}.$$

For generalized linear model a link function g satisfies that

$$g(u_{ij}) = \boldsymbol{x}_{ij}' \boldsymbol{\beta}.$$

Applying the chain rule to the matrix operations, we can obtain

$$\frac{\partial \boldsymbol{\mu}_i'}{\partial \boldsymbol{\beta}} \quad = \quad \begin{pmatrix} \frac{x_{i11}}{g'(u_{i1})} & \cdots & \frac{x_{in_i1}}{g'(u_{in_i})} \\ \vdots & \ddots & \vdots \\ \frac{x_{i1p}}{g'(u_{i1})} & \cdots & \frac{x_{in_ip}}{g'(u_{in_i})} \end{pmatrix}.$$

The covariance matrix for subject i is estimated by

$$\boldsymbol{V}_i \quad = \quad \boldsymbol{A}_i^{\frac{1}{2}} \boldsymbol{R}_i(\boldsymbol{\alpha}) \boldsymbol{A}_i^{\frac{1}{2}},$$

where $\boldsymbol{A}_i$ is an $n_i \times n_i$ diagonal matrix with $v(\mu_{ij})$ as the jth diagonal element, and $\boldsymbol{R}_i(\boldsymbol{\alpha})$ is an $n_i \times n_i$ working correlation matrix. It is estimated in the iterative fitting process

using the current value of the parameter vector $\boldsymbol{\beta}$ to compute appropriate functions of the Pearson residual as

$$e_{ij} = \frac{y_{ij} - \mu_{ij}}{\sqrt{v(\mu_{ij})}}.$$

The model-based covariance matrix of estimated $\boldsymbol{\beta}$ is

$$\text{Cov}(\hat{\boldsymbol{\beta}}) \quad = \quad \left(\sum_{i=1}^{K} \frac{\partial \boldsymbol{\mu}_i'}{\partial \boldsymbol{\beta}} \boldsymbol{V}_i^{-1} \frac{\partial \boldsymbol{\mu}_i}{\partial \boldsymbol{\beta}} \right)^{-1} \equiv \boldsymbol{V}_m,$$

which is same as the covariance estimate of the maximum likelihood estimator of $\boldsymbol{\beta}$ in generalized linear model.

The empirical, or robust estimator of the covariance matrix of $\boldsymbol{\beta}$ is

$$\text{Cov}(\hat{\boldsymbol{\beta}}) \quad = \quad \boldsymbol{V}_m^{-1} \left[\sum_{i=1}^{K} \frac{\partial \boldsymbol{\mu}_i'}{\partial \boldsymbol{\beta}} \boldsymbol{V}_i^{-1} \text{Cov}(\boldsymbol{Y}_i) \boldsymbol{V}_i^{-1} \frac{\partial \boldsymbol{\mu}_i}{\partial \boldsymbol{\beta}} \right] \boldsymbol{V}_m^{-1} \tag{3.5}$$

Note that Generalized Estimating Equations are not maximum likelihood functions. McCullagh (1983) proved that the estimates $\hat{\boldsymbol{\beta}}$ from GEEs converge to a normal distribution in large samples. The model-based covariance matrix of $\hat{\boldsymbol{\beta}}$ is only consistent if the mean model and the working correlation matrix are correct, while the empirical, or robust estimator of the covariance matrix of $\boldsymbol{\beta}$ is consistent even if the working correlation matrix is misspecified, i.e., $\text{Cov}(\boldsymbol{Y}_i) \neq \boldsymbol{V}_i$. Then the following estimate can be used to replace $\text{Cov}(\boldsymbol{Y}_i)$ in the (3.5)

$$\text{Cov}(\boldsymbol{Y}_i) \quad = \quad \left(\boldsymbol{Y}_i - \boldsymbol{\mu}_i(\hat{\boldsymbol{\beta}}) \right) \left(\boldsymbol{Y}_i - \boldsymbol{\mu}_i(\hat{\boldsymbol{\beta}}) \right)'.$$

3.2 The Models and Tests

3.2.1 Poisson Regression in a Log-linear Model

Poisson regression in the log-linear model is one of the most commonly used generalized linear models. It is applied to count data in which the response variable represents the number of events happening in a fixed period of time. To ensure the positive mean for

Poisson distribution, it is natural to assume that the logarithm of the expected mean is a linear function of explanatory variables. Suppose a random variable Y_i, $(i = 1, 2, \cdots n)$ follows a Poisson distribution. Let $\boldsymbol{x}_i = (x_{i1}, \cdots, x_{ip})$, and $\boldsymbol{\beta} = (\beta_1, \cdots, \beta_p)$ to be the vector of independent variables for the ith response and the vector of parameters as usual. We have a log-linear model such as

$$\log \mathrm{E}(Y_i) = \boldsymbol{x}_i'\boldsymbol{\beta}. \tag{3.6}$$

That is, the link function is log for Poisson regression. In this model the regression coefficient β_j represents the expected change in the log of the mean per unit change in the predictor x_{ij}. In other words increasing x_{ij} by one unit is associated with an increase of β_j in the log of the mean.

By exponentiating Equation (3.6) we obtain a multiplicative model for the mean itself:

$$\mathrm{E}(Y_i) \quad = \quad \exp(\boldsymbol{x}_i'\boldsymbol{\beta}).$$

In this model, an exponentiated regression coefficient $\exp(\beta_j)$ represents a multiplicative effect of the jth predictor on the mean. Increasing x_{ij} by one unit multiplies the mean by a factor $\exp(\beta_j)$

Note the mean and variance are same,

$$\mathrm{Var}(Y_i) = \mathrm{E}(Y_i) = \exp(\boldsymbol{x}_i'\boldsymbol{\beta}).$$

Thus, the usual assumption of homoscedasticity would not be appropriate for Poisson data since variance increase with mean.

The problem of lacking observations still exists when applying Poisson regression into sequence data. Sequence data are split by following the same technique as described in Chapter 2 to "increase" the number of replicates. We then have the following model,

$$Y_{ijk} \quad = \quad \mu + \alpha_i + \beta_j + \gamma_{ij} + \varepsilon_{ijk}, \tag{3.7}$$

for $i = 1, \cdots, a$; $j = 1, \cdots, b$; and $k = 1, \cdots, n$, where Y_{ijk} is the substitution rate for the

jth species at ith locus for replicate k; and α_i, β_j, and γ_{ij} represent gene effects, lineage effect and interacting between them respectively. The ε_{ijk}'s are assumed to be dependent, random variables with zero means and variances depending on the mean of response variable. Covariance can be modeled from data.

In order to apply GEEs to the log-linear model, we first need to identify the subject of interest, which will give the correlation structure of the data. We made the assumptions that substitution rates are independent across genes, but dependent within a gene at a particular replication. That is, the subject of our interest is substitution rates at a gene under the same replicate. Letting $\boldsymbol{y}_i$ be the ith subject, equation (3.7) can be written by matrix notation,

$$\boldsymbol{y}_i \quad = \quad \boldsymbol{x}_i\boldsymbol{\beta}.$$

The matrix $\boldsymbol{x}_i$ is a design matrix and the vector $\boldsymbol{\beta}$ is unknown coefficients. Note that $\boldsymbol{y}_i \sim \text{Poisson}\left(\exp(\boldsymbol{x}_i\boldsymbol{\beta}), \boldsymbol{V}_i\right)$, where $\boldsymbol{V}_i$ is the variance-covariance matrix in which variance depending on the mean of response variable and covariance will be modeled directly from data. After we obtain the estimates of $\boldsymbol{\beta}$ and covariance matrix, Wald statistics can be constructed to test the null hypothesis.

3.2.2 Wald Statistics

When our sample is large enough, we may obtain an approximation to the sampling distribution of the estimator $\boldsymbol{\beta}$ obtained by solving the GEEs as

$$\hat{\boldsymbol{\beta}} \sim \boldsymbol{N}(\boldsymbol{\beta}, \widehat{\boldsymbol{V}}_{\boldsymbol{\beta}}).$$

As in the ordinary generalized linear models, the Wald testing procedure is used to test null hypotheses of the form

$$\mathbf{H}_0 : \mathbf{L}\boldsymbol{\beta} \quad = \quad \mathbf{h},$$

where rank($\mathbf{L}$) is the raw number of $\mathbf{L}$. As usual, the large sample approximation is given

$$\mathbf{L}\hat{\boldsymbol{\beta}} \sim \boldsymbol{N}(\boldsymbol{L\beta}, \mathbf{L}\widehat{\mathbf{V}}_{\beta}\mathbf{L}').$$

Therefore, we may form the Wald-χ^2 statistics as follows,

$$\chi^2_* \quad = \quad (\mathbf{L}\hat{\boldsymbol{\beta}} - \mathbf{h})'(\mathbf{L}\widehat{\mathbf{V}}_{\boldsymbol{\beta}}\mathbf{L}')^{-\mathbf{1}}(\mathbf{L}\hat{\boldsymbol{\beta}} - \mathbf{h}).$$

which is approximately distributed as a χ^2 distribution with the degree of freedom equal to the number of rows of $\mathbf{L}$.

3.3 The Type I Errors of the Tests

3.3.1 The Parameter Sets

The primary simulations were conducted under the same parameter sets as the ANOVA-based methods to test the same null hypothesis as previous discussions. The way to test proportionality of branch lengths among loci is to test the interaction between lineage and locus effects. When we invoked PROC GENMOD from SAS software to get the estimated variance-covariance matrices, the procedure couldn't converge since singularity was present in the variance-covariance matrices in many cases. We tried different parameters and found that the procedure could converge well when the evolutionary distance between two homologous sequences is at least 0.1. Thus we increased the values of transitional and transversional rates in our simulations. The parameter sets used to investigate Type I errors for the null hypothesis at $\alpha = 0.05$ are illustrated in the following tables. As in the previous chapters, four sampling schemes are taken into account when investigating the Type I errors. The base frequencies used here are the same as those of ANOVA based-methods, and the transitional and transversional rates are increased to avoid the non-convergence in the GENMOD procedure. The following parameter sets represent different sampling schemes.

set 1: no interaction, no locus effects and no lineage effects

set 2: no interaction, having locus effects and no lineage effects

set 3: no interaction, no locus effects and having lineage effects

set 4: no interaction, having locus effects and lineage effect

Table 3.1: The first group of parameter sets for the GEE-based methods

Locus	Set	Base freq				Transition			Transversion		
		A	C	G	T	δ_1	δ_2	δ_3	ρ_1	ρ_2	ρ_3
A	1	.25	.25	.25	.25	.25	.25	.75	.25	.25	.75
	2	.25	.25	.25	.25	.25	.25	.50	.25	.25	.50
	3	.25	.25	.25	.25	.25	.75	1.0	.25	.75	1.0
	4	.25	.25	.25	.25	.10	.30	.45	.10	.30	.45
B	1	.25	.25	.25	.25	.25	.25	.75	.25	.25	.75
	2	.25	.25	.25	.25	.50	.50	1.0	.50	.50	1.0
	3	.25	.25	.25	.25	.25	.75	1.0	.25	.75	1.0
	4	.25	.25	.25	.25	.20	.60	.90	.20	.60	.90

Table 3.2: The second group of parameter sets for the GEE-based methods

Locus	Set	Base freq				Transition			Transversion		
		A	C	G	T	δ_1	δ_2	δ_3	ρ_1	ρ_2	ρ_3
A	1	.25	.25	.25	.25	.50	.50	.75	.25	.25	.375
	2	.25	.25	.25	.25	.25	.25	.50	.125	.125	.25
	3	.25	.25	.25	.25	.30	.50	.75	.15	.25	.375
	4	.25	.25	.25	.25	.20	.40	.50	.10	.20	.25
B	1	.25	.25	.25	.25	.50	.50	.75	.25	.25	.375
	2	.25	.25	.25	.25	.50	.50	1.00	.25	.25	.50
	3	.25	.25	.25	.25	.30	.50	.75	.15	.25	.375
	4	.25	.25	.25	.25	.40	.80	1.00	.20	.40	.50

Table 3.3: The third group of parameter sets for the GEE-based methods

Locus	Set	Base freq				Transition			Transversion		
		A	C	G	T	δ_1	δ_2	δ_3	ρ_1	ρ_2	ρ_3
A	1	.40	.10	.20	.30	.25	.25	.75	.25	.25	.75
	2	.40	.10	.20	.30	.25	.25	.50	.25	.25	.50
	3	.40	.10	.20	.30	.25	.75	1.0	.25	.75	1.0
	4	.40	.10	.20	.30	.10	.30	.45	.10	.30	.45
B	1	.40	.10	.20	.30	.25	.25	.75	.25	.25	.75
	2	.40	.10	.20	.30	.50	.50	1.0	.50	.50	1.0
	3	.40	.10	.20	.30	.25	.75	1.0	.25	.75	1.0
	4	.40	.10	.20	.30	.20	.60	.90	.20	.60	.90

Table 3.4: The fourth group of parameter sets for the GEE-based methods

Locus	Set	Base freq				Transition			Transversion		
		A	C	G	T	δ_1	δ_2	δ_3	ρ_1	ρ_2	ρ_3
A	1	.40	.10	.20	.30	.50	.50	.75	.25	.25	.375
	2	.40	.10	.20	.30	.25	.25	.50	.125	.125	.25
	3	.40	.10	.20	.30	.30	.50	.75	.15	.25	.375
	4	.40	.10	.20	.30	.20	.40	.50	.10	.20	.25
B	1	.40	.10	.20	.30	.50	.50	.75	.25	.25	.375
	2	.40	.10	.20	.30	.50	.50	1.00	.25	.25	.50
	3	.40	.10	.20	.30	.30	.50	.75	.15	.25	.375
	4	.40	.10	.20	.30	.40	.80	1.00	.20	.40	.50

3.3.2 The Simulation Results and Discussions

Simulations are conducted to compare the performance of the proposed ANOVA.L.C with GEE-based methods under both JC69 and HKY85 models. As introduced previously, both model-based and robust estimators of variance matrix can be obtained from GEEs. The terms GEE-model and GEE-robust will be used to represent methods by using corresponding variance estimators. Although we expect that the Wald statistics will be distributed as χ^2 distribution with one degree of freedom under the null hypothesis of proportionality of relative rates across loci, we wish to check this under different sequence lengths in order to establish a set of working rules. Two sets of 1000 replicate sets of three sequences with $500bp$, $1000bp$, and $10000bp$ each, respectively, are generated by creating two ancestral sequences by different base frequencies and then allowing the sequences to diverge independently by different combinations of transition and transversion rates. The descendent sequence data are then alternatively split into replicates 2, 5, and 10 for sequences with length $500bp$ and $1000bp$, and replicates 2, 5, 10, 20, and 40 for sequences with length $10000bp$, respectively. The substitution rates are estimated by the models JC69 and HKY85.

Simulations are conducted by invoking the GENMOD procedure from SAS Version 6.02 within my C++ program. The computational speed is also dependent on the total sequence length. The GENMOD procedure is about 300 times slower than that of the ANOVA-based method when we used SAS, or about twice as fast as the likelihood procedure. Simulation results for Type I errors at the 0.05 significance level are tabulated in the following tables. Note the numbers inside parentheses under the ANOVE.L.C are the percentage of non-positive substitution rates, while under the GEE-model are the sum of percentage of non-positive substitution rates and the non-convergence in GEE. The percentages under the GEE-robust are omitted since they are same as those under the GEE-model.

The simulation results from the ANOVA.L.C testing are fairly consistent with the previous simulations in which Type I errors had no significant changes as increasing number of replicates. In contrast, Type I errors for all sampling schemes reduce as the number of replicates increases in GEE-based methods. Look closely at simulation results from the

GEE-based methods, we see that Type I errors are in the range of 0.2 to 0.4 when the number of replicate is 2. Generally, most of Type I errors approach the desired level .05 when the number of replicates is at least 10. For the sequences with length $500bp$, 10 replicates will result in no solution in 30% to 70% cases. For the sequences with total length $1000bp$, most of Type I errors are around 0.06 to 0.08 which are still slightly away from the desired level 0.05. For the sequences with length $10000bp$, Type I errors are very close to the desired level as the number of replicates is increased to 40. Additionally, there are just a few cases without solution. This leads us to conclude that the GEEs-based method is more suitable to long sequence data with relatively big replicates than the short sequence. Furthermore, simulation results show that both the model-based and robust methods give similar Type I errors, which indicates that assumptions about correlation structure of substitution rates are probably fine. Comparing the results from both JC69 model and HKY85 model, we don't see any obvious difference even if the data are generated based on the HKY85 model.

Table 3.5: Comparing Type I errors (%) of the GEE-based methods with those of ANOVA.L.C for the sequences with length 500*bp* at the 1st group

Model JC69									
Set	ANOVA.L.C			GEE-model			GEE-robust		
	2	5	10	2	5	10	2	5	10
1	5.3	5.9(1.5)	6.3(20.0)	21.2	11.9(4.1)	6.3(31.7)	25.6	12.8	5.3
2	6.1	5.7(4.5)	5.0(30.3)	22.8	13.1(5.0)	7.2(36.7)	23.4	9.8	7.9
3	5.7	5.6(3.0)	5.2(41.2)	24.3	14.3(3.7)	7.1(44.4)	27.6	13.8	7.2
4	5.9	6.4(11.3)	4.7(47.6)	23.5	10.0 (16.2)	5.3(48.5)	27.5	11.2	4.2
Model HKY85									
1	5.0	5.6(1.5)	5.8(31.5)	26.7	15.3(5.9)	4.9(33.1)	27.0	14.4	4.8
2	6.4	6.2(5.6)	4.0(51.3)	24.4	11.3(8.1)	3.0(53.0)	24.7	11.9	3.2
3	5.6	6.5(8.7)	4.3(56.6)	25.7	12.0(10.7)	5.7(59.0)	25.3	10.3	4.1
4	5.7	5.3(16.2)	3.6(66.3)	26.1	12.8(25.1)	3.2(69.2)	26.1	12.9	3.7

Table 3.6: Comparing Type I errors (%) of the GEE-based methods with those of ANOVA.L.C for sequences with length 500*bp* at the 2nd group

Model JC69									
Set	ANOVA.L.C			GEE-model			GEE-robust		
	2	5	10	2	5	10	2	5	10
1	5.8	5.9(0.9)	5.8(28.2)	19.2	10.6(1.1)	6.3(33.7)	19.6	11.8	6.4
2	6.6	6.5(2.5)	6.0(38.5)	18.2	10.1(4.7)	7.2(39.9)	18.4	9.7	6.9
3	6.7	7.1(3.5)	5.2(42.2)	20.3	13.1(5.7)	9.1(44.9)	21.7	8.1	8.2
4	5.8	5.4(8.7)	5.3(47.5)	22.8	14.0(13.2)	4.3(50.5)	21.9	8.0	4.2
Model HKY85									
1	5.6	6.0(0.9)	5.8(31.5)	19.7	12.3(1.1)	5.9(34.1)	19.0	11.3	4.9
2	6.4	7.2(3.2)	6.4(42.3)	19.3	12.7(5.2)	6.3(43.0)	20.4	10.9	3.2
3	6.9	7.5(5.7)	4.3(56.1)	18.3	11.0(7.1)	8.7(60.0)	21.3	12.4	8.9
4	6.7	6.3(13.2)	3.6(60.3)	20.6	11.8(15.3)	3.2(68.9)	22.9	13.9	3.5

Table 3.7: Comparing Type I errors (%) of the GEE-based methods with those of ANOVA.L.C for sequences with length 500*bp* at the 3rd group

Model JC69									
Set	ANOVA.L.C			GEE-model			GEE-robust		
	2	5	10	2	5	10	2	5	10
1	6.3	6.9(0.9)	6.3(26.0)	25.2	12.6(2.1)	5.3(31.2)	25.5	12.8	5.5
2	7.1	6.5(6.5)	4.7(30.3)	24.2	11.1(9.0)	8.2(38.4)	23.4	9.2	7.3
3	6.3	7.1(6.7)	5.7(44.2)	26.3	13.1(9.7)	4.1(47.6)	27.6	13.1	4.2
4	5.9	6.4(12.3)	5.9(49.5)	28.5	16.5(14.2)	3.3(53.5)	27.5	12.8	3.4
Model HKY85									
1	5.3	6.0(0.7)	5.8(41.5)	26.7	13.3(1.9)	4.9(43.1)	27.0	14.0	4.9
2	7.4	7.2(5.0)	6.4(62.3)	24.3	12.7(5.2)	7.6(48.3.0)	24.3	11.9	7.0
3	6.9	7.5(5.7)	5.3(66.1)	25.3	13.0(5.7)	5.7(67.0)	25.8	12.4	5.2
4	6.7	6.3(15.2)	4.6(63.3)	26.9	13.8(15.9)	4.2(68.2)	26.6	12.9	3.4

Table 3.8: Comparing Type I errors (%) of the GEE-based methods with those of ANOVA.L.C for sequences with length 500*bp* at the 4th group

Model JC69									
Set	ANOVA.L.C			GEE-model			GEE-robust		
	2	5	10	2	5	10	2	5	10
1	6.1	6.5(0.5)	6.7(29.0)	25.4	12.6(1.1)	4.3(31.2)	25.6	13.8	4.3
2	6.0	6.7(3.5)	5.5(38.3)	25.1	12.1(4.0)	5.2(38.7)	23.4	10.2	4.9
3	6.7	7.7(3.0)	5.8(44.2)	26.4	13.5(3.7)	2.1(44.6)	25.6	12.1	2.2
4	5.4	5.4(10.3)	4.7(47.5)	27.0	13.8 (11.2)	4.3(47.5)	27.5	13.4	4.0
Model HKY85									
1	6.3	6.0(0.7)	5.4(41.5)	26.1	13.3(1.9)	4.9(43.1)	28.1	14.0	4.9
2	5.8	7.2(5.0)	4.4(62.3)	25.3	11.7(5.2)	3.0(63.0)	23.3	10.9	2.8
3	6.5	7.5(5.7)	3.3(66.1)	26.5	12.8(5.7)	3.7(67.0)	25.3	10.4	2.9
4	5.7	5.5(15.2)	3.7(63.3)	27.6	13.6(15.9)	3.2(68.2)	28.9	12.9	3.6

Table 3.9: Comparing Type I errors (%) of the GEE-based methods with those of ANOVA.L.C for sequences with length 1000*bp* at the 1st group

Set	ANOVA.L.C			GEE-model			GEE-robust		
	2	5	10	2	5	10	2	5	10
Model JC69									
1	5.2	5.8	5.3(0.0)	23.7	10.9	6.2(0.0)	21.5	9.7	6.4
2	6.3	6.4	6.6(2.8)	19.8	9.4	6.9(5.9)	19.2	9.9	6.6
3	6.1	5.9	6.3(14.1)	23.6	11.2	8.2(16.2)	23.9	11.6	8.7
4	6.4	6.0	5.5(6.1)	25.6	14.5	6.4(11.4)	25.7	13.4	7.0
Model HKY85									
1	5.7	5.9	5.5(2.5)	26.1	11.1	6.4(4.0)	25.7	11.4	6.0
2	6.4	6.8	5.9(6.0)	25.9	12.8	7.0(8.8)	26.5	11.5	6.7
3	5.7	5.8	6.4(15.5)	24.2	11.4	7.9(18.6)	24.5	11.9	8.2
4	6.1	6.6	6.5(9.3)	25.2	10.6	7.1(13.5)	27.5	13.4	7.6

Table 3.10: Comparing Type I errors (%) of the GEE-based methods with those of ANOVA.L.C for sequences with length 1000*bp* at the 2nd group

Set	ANOVA.L.C			GEE-model			GEE-robust		
	2	5	10	2	5	10	2	5	10
Model JC69									
1	5.6	5.8	5.6(1.0)	20.0	9.7	5.6(3.1)	20.5	8.7	4.0
2	5.9	6.3	6.0(4.7)	21.2	9.3	6.8(6.9)	22.2	9.8	7.0
3	6.0	6.9	6.3(12.1)	23.0	12.2	7.6(14.2)	23.4	11.6	8.1
4	6.4	7.0	6.5(10.1)	25.2	13.5	7.0(13.4)	25.2	13.3	6.9
Model HKY85									
1	6.0	6.5	6.5(2.6)	20.9	9.0	6.0(5.0)	20.7	10.6	6.1
2	6.4	6.8	6.4(6.0)	21.9	10.8	6.7(8.8)	22.9	12.5	5.6
3	6.7	6.8	6.4(14.2)	24.8	11.9	6.9(17.6)	24.9	11.6	6.5
4	7.0	7.2	6.8(19.4)	25.0	12.7	4.7(30.7)	27.1	13.4	4.4

Table 3.11: Comparing Type I errors (%) of the GEE-based methods with those of ANOVA.L.C for sequences with length 1000*bp* at the 3rd group

Model JC69									
Set	ANOVA.L.C			GEE-model			GEE-robust		
	2	5	10	2	5	10	2	5	10
1	5.8	6.6	6.0(1.0)	24.2	12.8	6.5(2.1)	25.1	12.7	6.6
2	6.3	6.4	5.6(2.7)	26.1	13.2	6.8(5.6)	26.4	13.0	6.9
3	5.7	6.0	6.1(4.8)	27.3	14.5	6.9(7.9)	26.2	10.8	6.2
4	6.8	6.9	6.3(11.2)	25.9	12.5	7.0(16.6)	26.0	11.7	5.8
Model HKY85									
1	5.4	6.1	7.5(3.0)	27.0	13.2	6.3(7.3)	27.8	12.9	6.7
2	6.5	7.0	6.4(7.6)	25.5	11.3	6.3(10.6)	26.3	10.6	5.9
3	6.0	6.8	7.0(10.3)	26.0	12.1	7.6(17.5)	26.9	12.7	6.1
4	7.3	7.1	6.5(16.2)	27.9	13.4	6.8(26.2)	27.8	13.1	6.9

Table 3.12: Comparing Type I errors (%) of the GEE-based methods with those of ANOVA.L.C for sequences with length 1000*bp* at the 4th group

Model JC69									
Set	ANOVA.L.C			GEE-model			GEE-robust		
	2	5	10	2	5	10	2	5	10
1	5.9	6.8	6.6(0.0)	23.2	12.7	6.4(0.0)	20.5	8.7	4.0
2	5.7	7.4	6.6(1.7)	19.2	9.3	5.8(1.9)	19.2	9.0	7.0
3	6.9	7.9	7.3(12.1)	23.3	12.2	7.8(14.2)	23.4	10.6	8.1
4	6.4	7.5	8.5(2.0)	25.4	13.5	6.0(30.4)	25.2	11.5	6.7
Model HKY85									
1	6.4	7.0	6.5(2.0)	23.9	10.2	6.0(2.0)	26.7	11.6	6.1
2	6.4	7.8	6.9(6.0)	20.9	9.8	6.7(6.8)	26.9	12.7	5.6
3	5.7	6.8	6.4(13.2)	24.8	11.9	6.9(17.6)	24.9	10.9	6.9
4	7.0	7.6	7.5(29.4)	26.0	10.7	5.3(39.7)	27.1	11.4	4.9

Table 3.13: Comparing Type I errors (%) of the GEE-based methods with those of ANOVA.L.C for sequences with length 10000*bp* at the 1st group

Set	ANOVA.L.C					GEE-model					GEE-robust				
	2	5	10	20	40	2	5	10	20	40	2	5	10	20	40
Model JC69															
1	5.9	5.8	5.7	5.0	5.6(0.0)	31.2	14.6	6.2	5.3	5.2(0.6)	29.3	13.1	5.6	5.5	5.1
2	5.6	5.5	5.3	5.4	5.6(0.0)	31.7	13.3	6.0	4.9	4.9(0.8)	31.4	13.4	5.4	5.0	5.0
3	6.0	5.7	6.1	5.5	6.1(0.0)	32.6	14.5	6.4	5.4	4.2(1.0)	32.8	14.7	6.8	5.8	4.4
4	5.6	5.3	5.2	5.0	5.4(0.7)	32.8	14.4	7.3	5.9	4.0(2.3)	34.2	15.2	7.3	6.5	5.6
Model HKY85															
1	5.1	5.7	5.2	5.6	5.6(0.0)	31.2	13.6	7.2	5.3	5.2(0.8)	30.2	13.2	7.0	5.8	5.2
2	5.5	5.3	5.5	5.9	5.4(0.0)	30.5	13.3	7.5	5.4.0	5.0(0.9)	32.6	14.1	7.4	6.0	5.4
3	5.6	6.3	6.7	6.0	6.5(0.0)	33.2	15.7	7.9	6.0	5.4(1.2)	34.6	16.1	7.9	6.0	5.3
4	5.8	6.1	5.4	6.1	5.9(3.9)	34.4	16.8	8.3	5.9	5.0(5.3)	33.6	13.8	7.2	5.6	4.3

Table 3.14: Comparing Type I errors (%) of the GEE-based methods with those of ANOVA.L.C for sequences with length 10000*bp* at the 2nd group

Set	ANOVA.L.C					GEE-model					GEE-robust				
	2	5	10	20	40	2	5	10	20	40	2	5	10	20	40
Model JC69															
1	6.0	6.8	6.4	7.0	7.6(0.0)	28.3	14.0	7.2	6.3	5.2(0.1)	29.3	15.1	7.6	7.0	6.1
2	5.9	6.0	7.2	6.4	7.0(0.0)	30.5	15.3	7.9	7.0	5.6(0.1)	30.7	14.6	7.0	6.7	5.6
3	6.0	7.0	7.1	6.6	6.1(0.0)	29.9	14.5	6.9	6.0	4.9(0.0)	30.8	15.7	7.8	6.8	5.4
4	6.1	6.7	7.2	7.0	7.4(0.9)	31.0	15.4	7.3	5.9	4.0(1.7)	29.2	14.2	8.3	6.5	5.5
Model HKY85															
1	6.1	6.7	6.7	6.6	6.6(0.0)	28.2	14.2	7.2	6.3	5.2(0.2)	30.2	15.2	8.0	7.0	6.2
2	5.8	6.3	6.0	6.4	6.5(0.0)	30.1	14.7	7.5	7.0	6.0(1.2)	30.1	14.5	7.4	7.0	5.4
3	6.3	6.7	6.8	7.0	6.5(0.6)	29.2	14.6	6.7	6.4	4.9(2.8)	29.6	15.1	6.9	6.0	5.3
4	5.8	6.1	6.4	6.1	6.2(3.3)	32.4	15.8	7.9	5.9	4.0(4.6)	28.7	14.8	7.2	5.8	4.3

Table 3.15: Comparing Type I errors (%) of the GEE-based methods with those of ANOVA.L.C for sequences with length 10000*bp* at the 3rd group

Set	ANOVA.L.C					GEE-model					GEE-robust				
	2	5	10	20	40	2	5	10	20	40	2	5	10	20	40
Model JC69															
1	6.4	7.0	7.0	6.8	6.6(0.0)	36.2	18.0	9.2	7.3	5.6(0.8)	37.0	18.1	9.6	7.5	5.8
2	5.9	6.0	7.2	6.4	7.0(0.0)	37.0	18.3	9.6.	7.6	5.7(0.5)	36.7	18.0	9.0	8.0	5.6
3	6.4	7.0	7.1	6.0	6.1(0.0)	38.9	19.5	9.9	8.0	5.9(0.0)	40.2	19.7	10.7	7.8	5.4
4	5.7	6.3	7.2	7.0	7.4(0.7)	40.6	20.0	10.3	7.9	5.0(4.3)	41.2	21.2	11.3	8.5	5.9
Model HKY85															
1	7.0	7.5	7.6	7.6	7.6(0.0)	37.1	17.2	9.0	7.3	5.2(1.2)	38.2	17.2	9.7	7.3	5.2
2	5.8	6.6	6.5	6.8	6.9(0.0)	36.5	16.3	9.5	7.5	6.0(1.1)	37.1	18.5	9.4	7.1	5.4
3	6.2	7.0	7.1	7.5	7.0(1.0)	39.2	18.6	9.9	7.8	6.0(1.2)	40.6	19.1	9.9	7.9	5.9
4	5.8	6.1	6.4	7.1	6.2(3.4)	40.5	19.8	9.3	6.9	5.4(4.9)	42.8	20.8	10.2	7.9	5.3

Table 3.16: Comparing Type I errors (%) of the GEE-based methods with those of ANOVA.L.C for sequences with length 10000*bp* at the 4th group

Set	ANOVA.L.C					GEE-model					GEE-robust				
	2	5	10	20	40	2	5	10	20	40	2	5	10	20	40
Model JC69															
1	6.9	7.8	7.7	7.0	6.6(0.0)	33.2	15.0	7.2	7.3	5.2(0.1)	33.3	15.1	7.6	7.5	6.1
2	5.9	6.0	7.2	6.4	6.0(0.0)	33.0	15.3	7.0	7.0	5.0(0.1)	34.7	16.0	8.0	8.0	5.0
3	6.0	6.0	6.1	6.0	6.1(0.0)	32.9	15.5	6.9	7.0	4.9(0.0)	32.8	15.7	7.8	7.8	5.4
4	6.1	7.3	7.2	7.0	6.4(0.7)	32.6	15.0	7.3	5.9	4.0(1.3)	33.2	16.2	8.3	7.5	5.9
Model HKY85															
1	7.1	7.7	7.0	7.6	7.6(0.0)	33.2	15.2	8.2	7.3	5.2(0.2)	33.2	14.2	8.0	7.3	5.2
2	5.8	6.3	6.5	6.0	6.9(0.0)	32.1	14.3	7.5	7.0	6.0(0.1)	32.1	14.5	7.4	7.0	5.4
3	6.6	7.3	7.0	7.0	7.5(0.0)	33.2	15.6	6.9	7.0	4.9(0.0)	33.6	15.1	6.9	7.0	5.9
4	5.8	8.1	6.4	7.1	6.2(3.4)	32.4	14.8	7.3	5.9	4.0(4.3)	32.8	14.8	7.2	5.9	4.3

3.4 The Powers of the Tests

We performed simulation to compare the power of GEE-based methods with ANOVA.L.C method under different phylogenetic trees. As mentioned before, there are some chances that no result is obtained if the evolutionary distance between two homologous sequences is small. So we should keep this in mind when we choose the parameters. From results of Type I errors, the GEE-based methods perform well on the long sequences with big replicates. Thus power study is conducted on the sequences with length $10000bp$. First two sets of ancestral sequences with length $10000bp$ are generated by different based frequencies, and then allowing the sequences to diverge along six paths, each with different substitution rates, independently by mutation. As shown in Table 3.17, the parameters used to investigate powers allow transition rates to be the twice of transversion rates. For the four parameter sets, besides for increasing branch length for species 2 at gene A, keep branch lengths for all the other species as a constant.

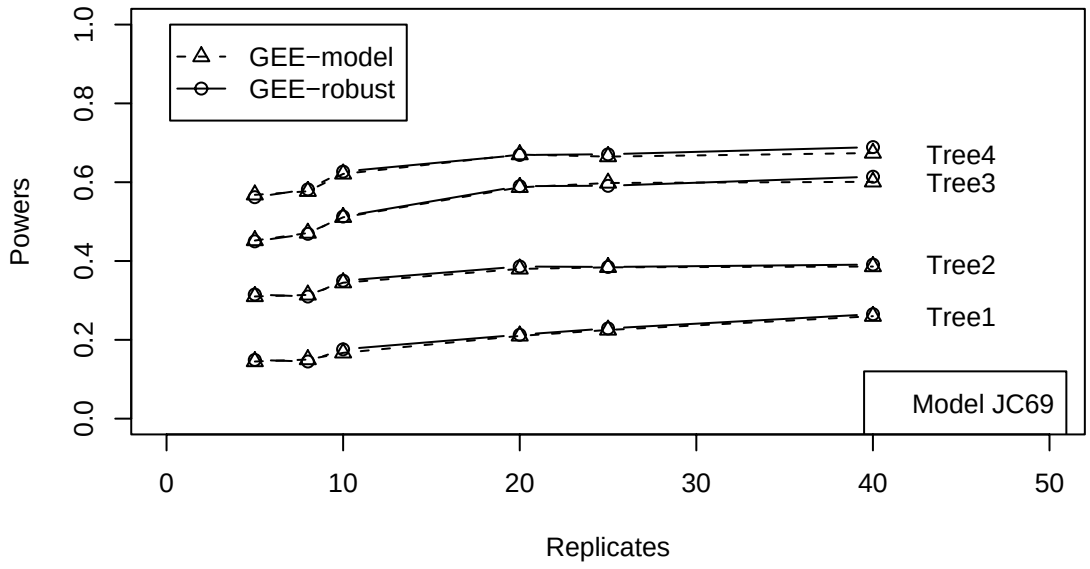
The results of power studies are shown in Figure 3.1 and 3.2. First, we compared the powers between methods based on GEE-model and GEE-robust under both the JC69 and HKY85 Models. As shown in Figure 3.1, the powers show the similar trend under the different phylogenetic trees for both GEE-model and GEE-robust methods as increasing the number of replicates. The magnitudes of powers under GEE-robust are slightly bigger than those under GEE-model. There are only two obvious increments of powers at the replicate 10 and 20 in almost all cases, and a slight increment of powers as increasing the number of replicates after 20. Comparing the powers under different phylogenetic trees, powers increases from about 15% to 70% as increasing the true differences from Tree 1 to Tree 4. Comparing the top panel with bottom panel in Figure 3.1, we can see that powers of the JC69 models are almost the same as those of the HKY85 model. This is quite consistent with the results from study of Type I errors.

Figure 3.2 shows the comparisons of powers between the GEE-based methods and ANOVA.L.C. at the replicates 20. The method based on GEE-model estimators was used to illustrate here. It is easy to see that the powers of ANOVA.L.C are bigger than those of

GEE-based methods. The magnitude of increase is about 50% under the same tree.

Table 3.17: The parameter sets of power study for the GEE-based methods

Locus	Set	Base freq				Transition			Transversion		
		A	C	G	T	δ_1	δ_2	δ_3	ρ_1	ρ_2	ρ_3
A	1	.40	.10	.20	.30	.20	.20	.50	.10	.10	.30
	2	.40	.10	.20	.30	.20	.20	.50	.10	.10	.30
	3	.40	.10	.20	.30	.20	.20	.50	.10	.10	.30
	4	.40	.10	.20	.30	.20	.20	.50	.10	.10	.30
B	1	.40	.10	.20	.30	.20	.24	.50	.10	.12	.30
	2	.40	.10	.20	.30	.20	.26	.50	.10	.13	.30
	3	.40	.10	.20	.30	.20	.28	.50	.10	.14	.30
	4	.40	.10	.20	.30	.20	.30	.50	.40	.15	.30



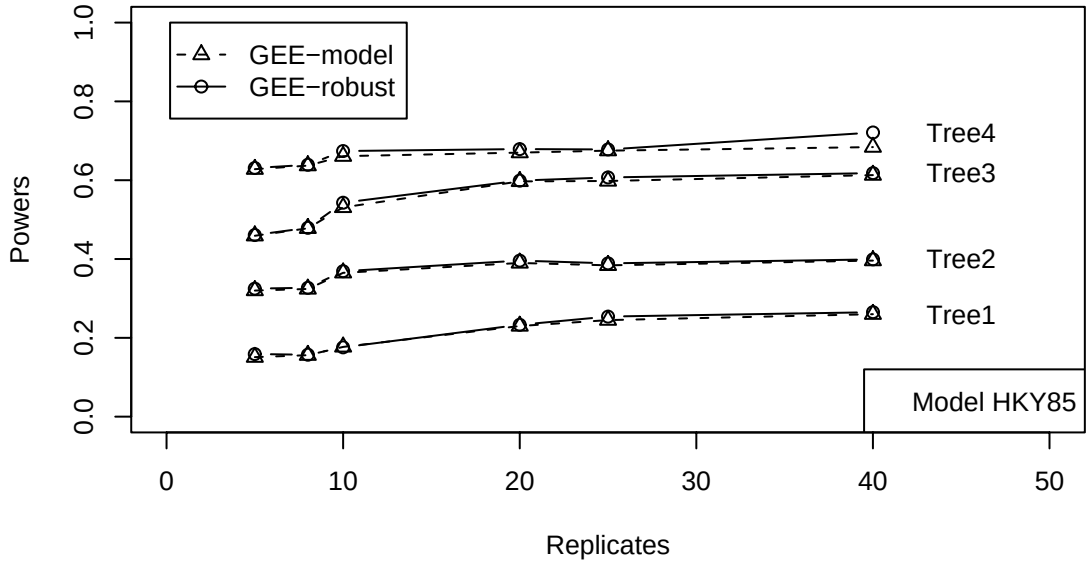
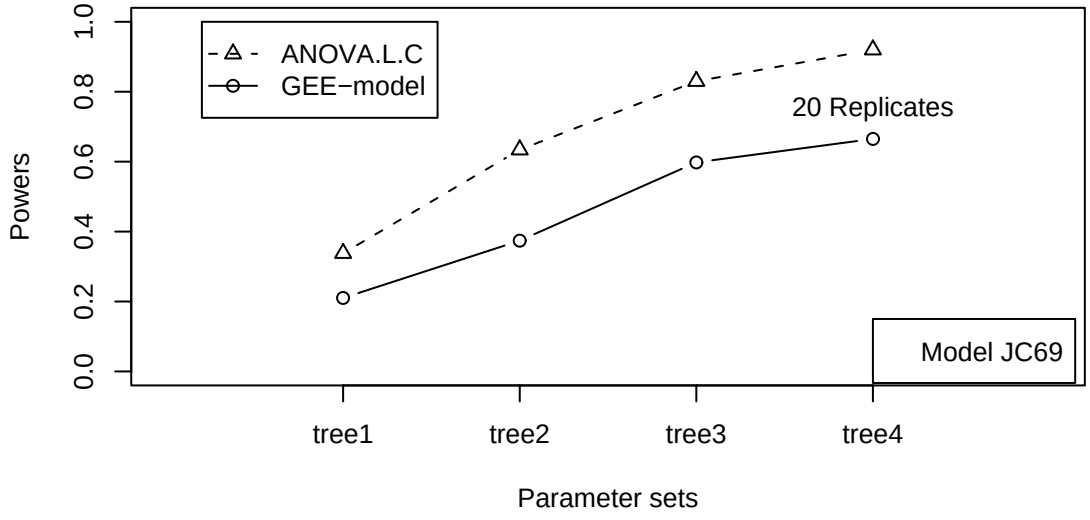


Figure 3.1: Power (%) study of GEE-based methods



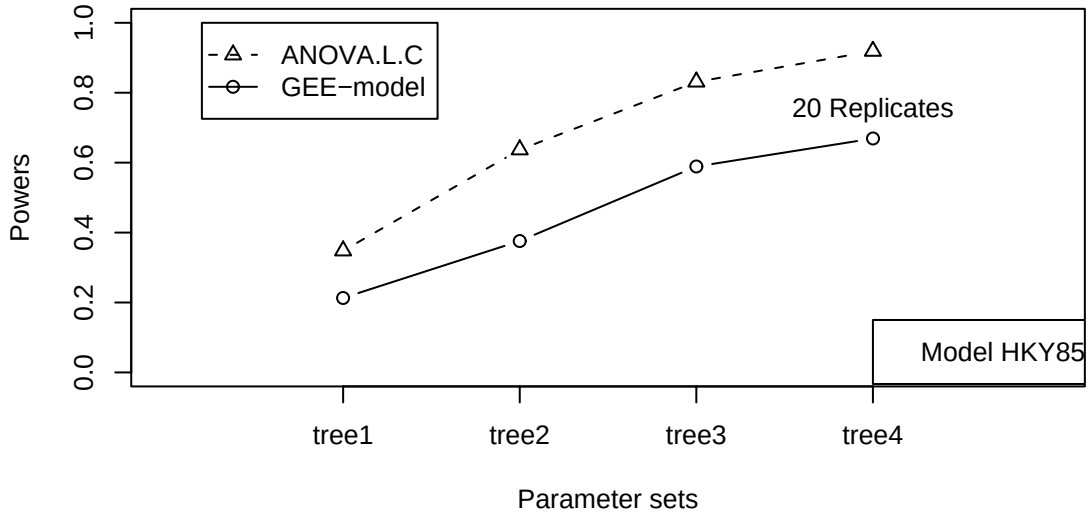


Figure 3.2: Comparing the powers (%) of GEE-based methods with those of ANOVA.L.C

Chapter 4

The Methods Based on Bootstrap Percentile Confidence Intervals

4.1 Introduction

4.1.1 Relation between Confidence Intervals and Hypothesis Tests

To better understand the proposed methods, we first review the basic relation between confidence intervals and hypothesis tests. Suppose we have an estimator $\hat{\beta}$ which is believed to be normally distributed as:

$$\hat{\beta} \sim N(\beta, s^2),$$

with the known standard error s. The following random variable then has a standard normal distribution,

$$Z = \frac{\hat{\beta} - \beta}{s} \sim N(0, 1).$$

The $(1 - 2\alpha)$ confidence interval of $(\hat{\beta} - \beta)/s$ is

$$z^{(\alpha)} \leq \frac{\hat{\beta} - \beta}{s} \leq z^{(1-\alpha)}.$$

Thus, the $(1 - 2\alpha)$ confidence interval of β is given by

$$[\hat{\beta} + sz^{(\alpha)}, \quad \hat{\beta} + sz^{(1-\alpha)}].$$

Let $\hat{\beta}_{\text{lo}} = \hat{\beta} + sz^{(\alpha)}$ and $\hat{\beta}_{\text{up}} = \hat{\beta} + sz^{(1-\alpha)}$. The probability that the interval $[\hat{\beta}_{\text{lo}}, \hat{\beta}_{\text{up}}]$ contains true value of β is $1 - 2\alpha$. Usually, the probability that β is below the lower bound is exactly α, as is the probability that β is above the upper bound,

$$\text{Prob}_\beta\{\beta < \hat{\beta}_{\text{lo}}\} = \alpha, \qquad \text{Prob}_\beta\{\beta > \hat{\beta}_{\text{up}}\} = \alpha.$$

From the hypothesis testing point of view, we have another way to understand the $1 - 2\alpha$ confidence interval $[\hat{\beta}_{\text{lo}}, \hat{\beta}_{\text{up}}]$ for β. Suppose the true $\beta = \hat{\beta}_{\text{lo}}$ and let the related test statistics be β^*, and

$$\beta^* \sim N(\hat{\beta}_{\text{lo}}, s^2).$$

It is obvious that the probability that β^* exceeds the estimate $\hat{\beta}$ is α,

$$\text{Prob}_\beta\{\beta^* \geq \hat{\beta}\} = \alpha.$$

Then for any true value of β smaller than the lower bound $\hat{\beta}_{lo}$, we can have

$$\text{Prob}_\beta\{\beta^* \geq \hat{\beta}\} < \alpha. \tag{4.1}$$

Likewise, for any true value of β larger than the upper bound $\hat{\beta}_{up}$, we can have

$$\text{Prob}_\beta\{\beta^* \leq \hat{\beta}\} < \alpha. \tag{4.2}$$

Now we can say that the $1 - 2\alpha$ confidence interval $[\hat{\beta}_{\text{lo}}, \hat{\beta}_{\text{up}}]$ for β is the set of values of β with observed $\hat{\beta}$, and we can't exclude those values by both of the tests (4.1) and (4.2). That is, a hypothesis test can be performed by constructing a confidence interval and then checking whether the true value is in the interval or not. If the confidence interval fails to include the true value, then the test is deemed to be rejected at the significance level 2α.

4.1.2 Confidence Intervals Based on Bootstrap Percentile

The bootstrap is a computer-based resampling method for estimating some statistical parameters of interest. To illustrate the idea of bootstrap procedure, we consider the following

case: we have a random sample $\mathbf{x} = (x_1, x_2, \cdots, x_n)$ from an unknown probability distribution F, and we want to estimate a parameter of interest $\beta = g(\mathbf{x})$. Given that we have no other information about the population, the sample $\mathbf{x}$ is the single best estimate of that population. We therefore treat that sample as the population and use bootstrap sampling to generate a series of resamples from the original sample. Suppose $\hat{F}$ is the empirical distribution of F, and a bootstrap sample is a random sample of size n drawn from $\hat{F}$ with replacement, say $\mathbf{x}^* = (x_1^*, x_2^*, \cdots, x_n^*)$. That is, each data point in the sample $\mathbf{x}$ has equal probability $1/n$ to be drawn. The one of advantages of bootstrap is that one can draw any number of resample that one wants. Suppose we draw B bootstrap samples $(\mathbf{x}_1^*, \mathbf{x}_2^*, \cdots, \mathbf{x}_B^*)$, then we can obtain $\hat{\boldsymbol{\beta}}^* = (\hat{\beta}_1^*, \hat{\beta}_2^*, \cdots, \hat{\beta}_B^*)$ from B bootstrap samples,

$$\hat{\beta}_B^* = g(\mathbf{x}_B^*).$$

Suppose $\hat{F}^*$ is the cumulative distribution function of $\hat{\boldsymbol{\beta}}^*$. The $1-2\alpha$ percentile interval is given by the α and $1-\alpha$ percentiles of $\hat{F}^*$:

$$[\hat{\beta}_{\text{lo}}, \hat{\beta}_{\text{up}}] \quad = \quad [\hat{F}^{*-1}(\alpha), \hat{F}^{*-1}(1-\alpha)].$$

4.2 The Methods and Simulations

4.2.1 Estimate of Parameters

Let Y_{ijk} be the substitution rate for the jth species at the ith locus within the kth replicate. Consider the following two-way ANOVA model

$$Y_{ijk} \quad = \quad \mu + \alpha_i + \beta_j + \gamma_{ij} + \varepsilon_{ijk}, \tag{4.3}$$

for $i = 1, \cdots, a$; $j = 1, \cdots, b$; and $k = 1, \cdots, n$. Where μ denotes the grand mean, α_i denotes the ith locus effect, β_j denotes the jth lineage effect, γ_{ij} denotes the interaction between the jth lineage and the ith locus effects, and ε_{ijk} are independent $N(0, \sigma^2)$.

To obtain the least squares estimators, we minimize the following expression:

$$Q = \sum_i \sum_j \sum_k [Y_{ijk} - \mu - \alpha_i - \beta_j - \gamma_{ij}]^2$$

under the restrictions:

$$\sum_{i=1}^{a} \alpha_i = 0, \qquad \sum_{j=1}^{b} \beta_j = 0, \qquad \sum_{i=1}^{a} \gamma_{ij} = 0, \qquad \sum_{j=1}^{b} \gamma_{ij} = 0.$$

We obtain the following least squares estimators of the parameters:

$$\hat{\mu} = \bar{Y}_{...}, \quad \hat{\alpha}_i = \bar{Y}_{i..} - \bar{Y}_{...}, \quad \hat{\beta}_i = \bar{Y}_{.j.} - \bar{Y}_{...}, \quad \hat{\gamma}_{ij} = \bar{Y}_{ij.} - \bar{Y}_{i..} - \bar{Y}_{.j.} + \bar{Y}_{...} \tag{4.4}$$

Note that those estimators can still be obtained if the number of replicate $k = 1$.

Similarly, those least squares estimators also can be obtained in the same way for the following model:

$$\log(Y_{ijk}) = \mu + \alpha_i + \beta_j + \gamma_{ij} + \varepsilon_{ijk}, \tag{4.5}$$

4.2.2 Hypothesis Tests Based on Confidence Intervals

The way to test proportionality of branch lengths among loci is to test the interaction between lineage and locus effects. That is, the hypotheses of the proposed method based on the model (4.3) are defined as

$$H_0 : \text{all} \quad \gamma_{ij} = 0, \qquad H_1 : \text{at least one of} \quad \gamma_{ij} \neq 0. \tag{4.6}$$

Suppose the $1 - 2\alpha$ confidence intervals for the estimate $\hat{\gamma}_{ij}$ are $[\hat{\gamma}_{ij}^{\text{lo}}, \hat{\gamma}_{ij}^{\text{up}}]$ which will be constructed by using bootstrap resampling. The test (4.6) is equivalent to checking whether the confidence intervals $[\hat{\gamma}_{ij}^{\text{lo}}, \hat{\gamma}_{ij}^{\text{up}}]$ contain the null value 0. If $[\hat{\gamma}_{ij}^{\text{lo}}, \hat{\gamma}_{ij}^{\text{up}}]$ excludes the null value 0, then we can reject the null hypothesis at the significance level α.

4.2.3 The Type I Errors of the Tests

There are two kinds of tests based on bootstrap percentile confidence intervals. We name one based on Model (4.3) as BOOTPCI, while calling another one based on Model (4.5)

as BOOTPCI.L. We use the same parameter sets as those in ANOVA-based methods in chapter2 to generate the original DNA sequences for three species at two loci, respectively. The study of Type I errors is conducted on sequences with lengths $500bp$ and $1000bp$, respectively. 1000 bootstrap samples of three species at two loci are then drawn column by column for three species within each locus, separately, from that original DNA sequences. That is, we treat the three nucleotides at the same position for the three species at the same gene as a data point. Thus, we draw them together and put them in the same position to the corresponding DNA sequences. Although we do not have the problem of lacking replicate like all the previously proposed methods here, the sequences are still alternatively split as before in order to compare with other proposed methods.

We compute 1000 bootstrap estimates $\hat{\boldsymbol{\gamma}}^*_{ij} = (\hat{\gamma}^{*1}_{ij}, \hat{\gamma}^{*2}_{ij}, \cdots, \hat{\gamma}^{*1000}_{ij})$ according to the formula (4.4) from the 1000 bootstrap samples. We then sort the 1000 $\hat{\boldsymbol{\gamma}}^*_{ij}$ in ascending order, and obtain values of $\hat{\gamma}^*_{ij}$ at the positions 25 and 975, denoting as $\hat{\gamma}^*_{ij}(25)$ and $\hat{\gamma}^*_{ij}(975)$. Thus, the 95% percentile confidence intervals based on bootstrap are $[\hat{\gamma}^*_{ij}(25), \hat{\gamma}^*_{ij}(975)]$. Comparison of the null value 0 of $\hat{\gamma_{ij}}$ with the intervals $[\hat{\gamma}^*_{ij}(25), \hat{\gamma}^*_{ij}(975)]$ will be conducted to check whether the interval contains 0. The total 1000 such simulations are conducted in the way as described. Let m denote the total number of simulations that the 95% confidence intervals fail to contain the value 0 in the 1000 times of simulation, the Type I error can be calculated by $m/1000$.

The Simulation study is conducted to compare Type I errors for LRT, ANOVA.L.C and Bootstrap versions of the relative ratio test under the model of JC69 for the sequences with lengths $500bp$ and $1000bp$, respectively. For convenience, we put the simulation results from the lengths $500bp$ and $1000bp$ at the same parameter set into one table. First, we examine the Type I errors from the bootstrap based methods. Generally, BOOTPCI.L, the method with taking logarithms of substitution rates, gives smaller Type I errors than does the BOOTPCI method. In particular, for the 4th sampling scheme, having both lineage and locus effects, the BOOTPCI.L performs much better than the BOOTPCI method in which the Type I errors are in the range of .15 to .40. For the four sets of parameters, Type I errors

produced by the BOOTPCI.L method are very close to 0.05 at the significance level $\alpha = .05$ for most cases. As increasing the number of replicates to 5, most Type I errors from the BOOTPCI.L method are closer to the desired value of 0.05 than those from simulations with 1 replicate. Note that some of Type I errors are very small for the sequences with length $500bp$ at the replicate 5. The reason is that a certain amount of estimated substitution rates are not positive values when we split the sequence with length $500bp$ into 5 pieces of sequences. For the simulation using bootstrap resampling, it is impossible to count that proportions like the ANOVA.L.C method and include them in the tables. By checking the values of percentages inside the parentheses for the ANOVA.L.C method, we can have a clear picture of the 1000 original data for bootstrap. The larger the percentages from the ANOVA.L.C method are, the more unreliable Type I errors from the BOOTPCI.L method.

We compare the Type I errors from the BOOTPCI.L with both methods of ANOVA.L.C and LRT. It is not surprising to see that BOOTPCI.L results in smaller Type I errors than does ANOVA.L.C method in 95% of the cases. In contrast, it is surprising that the BOOTPCI.L gives more desirable Type I errors than does LRT method in general. We compare the speed of computing between LRT-based method with the BOOTPCI.L. At one simulation time, it takes 12 seconds to obtain the result from the BOOTPCI.L method, and 1 second from the LRT based method for the sequence length with $1000bp$ under the SUNBlade 1000's with 750Mhz processors with 1 GB of memory. That is the speed of BOOTPCI.L is about 12 times slower than that of LRT based method.

Table 4.1: Comparing Type I errors (%) of the bootstrap-based methods with those of ANOVA.L.C at the 1st group

Sequence Length 500*bp*									
Set	LRT	ANOVA.L.C		BOOTPCI			BOOTPCI.L		
		2	5	1	2	5	1	2	5
1	5.2	6.9	5.4(3.0)	4.3	4.4	4.3	4.1	4.2	3.3
2	5.3	6.7	6.0(3.7)	6.1	6.2	5.6	5.3	5.1	4.1
3	5.4	7.6	6.9(22.7)	4.1	4.0	4.1	3.9	3.4	1.8
4	5.6	6.6	5.4(26.0)	18.2	18.3	18.6	4.3	4.3	2.9
Sequence Length 1000*bp*									
1	5.0	4.4	5.6(0.0)	6.0	6.1	6.1	5.9	5.8	5.4
2	4.8	6.5	6.4(0.0)	5.9	5.6	5.7	5.5	5.6	5.4
3	5.2	6.3	6.8(0.0	4.6	4.9	4.6	4.7	4.8	4.0
4	4.4	5.2	5.2(0.0)	28.0	28.0	28.0	5.1	4.9	4.1

Table 4.2: Comparing Type I errors (%) of the bootstrap-based methods with those of ANOVA.L.C at the 2nd group

Sequence Length 500*bp*									
Set	LRT	ANOVA.L.C		BOOTPCI			BOOTPCI.L		
		2	5	1	2	5	1	2	5
1	5.1	5.4	5.2(1.3)	4.2	4.4	4.0	4.1	4.3	2.9
2	5.9	6.9	7.6(2.4)	6.3	6.2	6.0	6.0	6.0	5.9
3	5.7	6.6	6.2(2.7)	6.0	5.9	5.8	4.9	4.7	3.3
4	4.6	6.4	5.6(8.8)	23.7	23.6	23.7	5.9	5.7	3.5
Sequence Length 1000*bp*									
1	4.2	5.7	5.0(0.0)	5.4	5.1	5.3	5.4	5.6	4.8
2	5.3	5.9	6.4(0.0)	5.0	5.6	5.0	5.3	4.9	4.8
3	4.4	5.0	5.5(0.0)	4.4	4.3	4.5	4.9	4.8	3.9
4	5.0	5.0	5.1(0.0)	38.9	39.1	39.4	5.5	5.3	5.1

Note: the numbers inside the parentheses are percentages of non-positive branch lengths for method of ANOVA.L.C.

Table 4.3: Comparing Type I errors (%) of the bootstrap-based methods with those of ANOVA.L.C at the 3rd group

Sequence Length 500*bp*									
Set	LRT	ANOVA.L.C		BOOTPCI			BOOTPCI.L		
		2	5	1	2	5	1	2	5
1	5.1	4.4	4.5(47.1)	4.5	4.4	4.3	4.4	4.3	3.2
2	6.0	5.1	6.0(28.0)	6.2	5.7	5.8	5.7	5.4	4.1
3	4.8	5.5	5.5(32.2)	5.3	5.2	5.1	5.2	4.0	2.2
4	5.0	7.1	4.6(28.9)	16.3	15.7	15.5	5.1	4.3	2.2
Sequence Length 1000*bp*									
1	6.2	5.2	5.8(1.6)	5.9	5.6	5.9	5.5	5.1	4.0
2	5.4	5.1	5.5(1.3)	6.0	6.0	5.8	5.3	5.0	4.3
3	4.7	6.7	5.8(4.5)	5.0	5.1	4.8	4.6	4.5	3.4
4	6.3	7.3	7.0(2.9)	25.5	25.9	25.2	6.1	6.0	4.0

Table 4.4: Comparing Type I errors (%) of the bootstrap-based methods with those of ANOVA.L.C at the 4th group

Sequence Length 500*bp*									
Set	LRT	ANOVA.L.C		BOOTPCI			BOOTPCI.L		
		2	5	1	2	5	1	2	5
1	5.3	6.1	4.4(28.7)	4.8	4.7	4.6	5.3	4.5	2.1
2	5.6	5.8	5.6(10.6)	4.5	4.3	4.3	4.9	4.5	3.0
3	4.1	5.8	4.6(28.1)	5.0	4.9	4.7	4.8	4.4	2.5
4	5.9	6.2	5.7(8.9)	18.9	18.9	18.8	6.7	6.6	4.9
Sequence Length 1000*bp*									
1	5.6	6.7	6.2(1.5)	5.2	5.1	5.1	4.9	4.9	3.8
2	5.8	6.1	6.8(0.0)	4.5	4.4	4.6	5.8	5.6	5.6
3	4.9	5.2	6.7(3.2)	5.4	5.4	5.4	5.6	5.1	4.6
4	6.4	6.6	7.3(0.0)	31.8	31.6	31.6	6.1	5.9	5.4

4.2.4 The Power of the Tests

Simulations are also conducted to compare statistical powers of BOOTBCI.L method with ANOVA.L.C and LRT based methods. We use the same parameter sets as those in the power study of ANOVA-based methods. It is shown in Table 2.19. The parameters used to investigate powers allow transition rates to be twice of transversion rates and unequal base frequencies. For the four parameter sets, besides increasing the branch lengths for species 2 at gene A accordingly, we keep branch lengths for all the other species as constant in the meantime. Sequences with length $2000bp$ are used in primary simulations. The numbers of replicates is $2, 5, 8, 10$ for the ANOVA.L.C, and $1, 2, 5, 8, 10$ for the BOOTPCI.L. The JC69 model is used, and the significance level $\alpha = .05$ as usual.

In order to easily see the subtle differences of power among different methods, we illustrate the results in Table 4.2.4. We surprisingly see that the powers for both LRT and ANOVA.L.C are much smaller than those from previous investigation of ANOVA based methods in which we used the sequences with length $10000bp$. For example, the powers of ANOVA.L.C and LRT are around 40% at Tree 4 here, in contrast, those are close to 100% if using the sequences with length $10000bp$. It is interesting to see that powers increase as the number of replicates increase for the ANOVA.L.C method, and the powers slightly decrease as the number of replicates is increasing for the BOOTPCI.L method. The greatest powers for both BOOTPCI.L and ANOVA.L.C method are pretty close to each other, and also close to the powers of LRT method.

Table 4.5: Comparing powers (%) of BOOTPCI.L with those of both LRT and ANOVA.L.C methods

Tree	LRT	ANOVA.L.C				BOOTPCI.L				
		2	5	8	10	1	2	5	8	10
1	12.5	9.90	12.5	13.0	13.9	12.5	12.2	11.8	11.0	10.6
2	19.9	14.8	18.6	18.8	19.6	19.9	20.1	18.7	17.2	16.8
3	28.2	20.7	27.1	27.6	27.5	27.1	27.4	26.2	24.6	24.1
4	41.0	27.1	40.0	39.9	40.3	40.6	40.9	39.9	38.0	37.0

Chapter 5

The Methods Based on Weighted Least Squares Estimation with Covariance Structure

5.1 Introduction

In Chapter 2 we introduced a two-way ANOVA model with nonzero covariances. Although the proposed method ANOVA.L.C has better power than LRT-based relative ratio test, the Type I errors are slightly away from the desired level. We suspect that the assumption of homogenous error variances may be the major reason, so we went back to examine our simulated data of substitution rates and found the following interesting results. The sample variances of substitution rates are approximately a linear function of sample means with a constant coefficient. The sample covariances between two substitution rates at the same gene are approximately a linear function of the product of the square root of the two sample means.

Based on this observation, we assume that covariances between two substitution rates at the same gene are a linear function of the product of the square root of the means. We propose a new method which uses iterative weighted least squares (IWLS) estimation with variance and covariance structure. The typical IWLS procedure just deals with heteroscedastic linear regression. The challenge here is how to incorporate covariance structure into the weight also. We will derive the formula here. The implementation will be the future work.

5.2 IWLS with Covariance Structure

5.2.1 Deriving the Formulas of Estimating Parameters

Based on previous introduction, we assume that

$$Y_{ijk} = \mu + \alpha_i + \beta_j + \gamma_{ij} + \varepsilon_{ijk}$$

for $i = 1, \cdots, a$; $j = 1, \cdots, b$; and $k = 1, \cdots, n$, where α_i, β_j, and γ_{ij} represent gene effects, lineage effects and interaction between them, respectively. For simplicity of discussion, assume that $b = 2$. If $b > 2$, the idea is similar but more complicated.

Further assume that f_{ij}, $i = 1, \cdots, a$; $j = 1, \cdots, b$ are positive known constants.

Assume that

$$\mathrm{Cov}(\varepsilon_{i_1,j_1,k_1}, \varepsilon_{i_2,j_2,k_2}) = \begin{cases} 0, & \text{if } i_1 \neq i_2 \\ 0, & \text{if } k_1 \neq k_2 \\ \sigma^2 f_{i_1,j_1}, & \text{if } i_1 = i_2, j_1 = j_2, \text{ and } k_1 = k_2 \\ \sigma_{i12}\sqrt{f_{i_1,j_1} f_{i_2,j_2}}, & \text{if } i_1 = i_2, j_1 \neq j_2, \text{ and } k_1 = k_2 \end{cases}$$

and $E(\varepsilon_{ijk}) = 0$.

If f_{ij}'s are constant 1 for $i = 1, \cdots, a$; $j = 1, \cdots, b$, this model will become the model we discussed before.

Make the following transformation:

$$\begin{aligned} Y^{\star}_{ijk} &= Y_{ijk}/\sqrt{f_{ij}}, \\ \varepsilon^{\star}_{ijk} &= \varepsilon_{ijk}/\sqrt{f_{ij}}. \end{aligned}$$

Then the transformed model is

$$Y^{\star}_{ijk} = \mu/\sqrt{f_{ij}} + \alpha_i/\sqrt{f_{ij}} + \beta_j/\sqrt{f_{ij}} + \gamma_{ij}/\sqrt{f_{ij}} + \varepsilon^{\star}_{ijk},$$

and

$$\mathrm{Cov}(\varepsilon^{\star}_{i_1,j_1,k_1}, \varepsilon^{\star}_{i_2,j_2,k_2}) = \begin{cases} 0, & \text{if } i_1 \neq i_2 \\ 0, & \text{if } k_1 \neq k_2 \\ \sigma^2, & \text{if } i_1 = i_2, j_1 = j_2, \text{ and } k_1 = k_2 \\ \sigma_{i12}, & \text{if } i_1 = i_2, j_1 \neq j_2, \text{ and } k_1 = k_2 \end{cases}$$

Assume that $\sigma_{i12} < 0$ and $\sigma^2 > -\sigma_{i12}$. Moreover, assume that σ^2 and σ_{i12} are known constants for $i = 1, \cdots, a$.

Let

$$A_i = \begin{bmatrix} \sigma^2 & \sigma_{i12} \\ \sigma_{i12} & \sigma^2 \end{bmatrix}.$$

It is easy to verify that A_i's two eigenvalues are $\sigma^2 - \sigma_{i12}$ and $\sigma^2 + \sigma_{i12}$ and their eigenvectors are

$$\begin{pmatrix} 1 \\ -1 \end{pmatrix}, \quad \text{and} \quad \begin{pmatrix} 1 \\ 1 \end{pmatrix}.$$

respectively.

Therefore,

$$A_i = \begin{bmatrix} 1/\sqrt{2} & 1/\sqrt{2} \\ -1/\sqrt{2} & 1/\sqrt{2} \end{bmatrix} \begin{bmatrix} \sigma^2 - \sigma_{i12} & \\ & \sigma^2 + \sigma_{i12} \end{bmatrix} \begin{bmatrix} 1/\sqrt{2} & -1/\sqrt{2} \\ 1/\sqrt{2} & 1/\sqrt{2} \end{bmatrix}.$$

This leads to making the following transformation:

$$\begin{aligned} Y^{\star\star}_{i1k} &= \frac{1}{\sqrt{2(\sigma^2 - \sigma_{i12})}}(Y^{\star}_{i1k} - Y^{\star}_{i2k}), \\ Y^{\star\star}_{i2k} &= \frac{1}{\sqrt{2(\sigma^2 + \sigma_{i12})}}(Y^{\star}_{i1k} + Y^{\star}_{i2k}). \end{aligned}$$

Or equivalently

$$\begin{aligned} Y^{\star\star}_{i1k} &= \frac{1}{\sqrt{2(\sigma^2-\sigma_{i12})}}\left(\frac{Y_{i1k}}{\sqrt{f_{i1}}}-\frac{Y_{i2k}}{\sqrt{f_{i2}}}\right), \\ Y^{\star\star}_{i2k} &= \frac{1}{\sqrt{2(\sigma^2+\sigma_{i12})}}\left(\frac{Y_{i1k}}{\sqrt{f_{i1}}}+\frac{Y_{i2k}}{\sqrt{f_{i2}}}\right). \end{aligned}$$

Similarly,

$$\begin{aligned} \varepsilon^{\star\star}_{i1k} &= \frac{1}{\sqrt{2(\sigma^2-\sigma_{i12})}}\left(\frac{\varepsilon_{i1k}}{\sqrt{f_{i1}}}-\frac{\varepsilon_{i2k}}{\sqrt{f_{i2}}}\right), \\ \varepsilon^{\star\star}_{i2k} &= \frac{1}{\sqrt{2(\sigma^2+\sigma_{i12})}}\left(\frac{\varepsilon_{i1k}}{\sqrt{f_{i1}}}+\frac{\varepsilon_{i2k}}{\sqrt{f_{i2}}}\right). \end{aligned}$$

That is,

$$\begin{aligned} Y^{\star\star}_{i1k} &= \frac{1}{\sqrt{2(\sigma^2-\sigma_{i12})}}\left[(\frac{1}{\sqrt{f_{i1}}}-\frac{1}{\sqrt{f_{i2}}})\mu+(\frac{1}{\sqrt{f_{i1}}}-\frac{1}{\sqrt{f_{i2}}})\alpha_i\right. \\ &\quad \left.+\frac{\beta_1}{\sqrt{f_{i1}}}-\frac{\beta_2}{\sqrt{f_{i2}}}+\frac{\gamma_{i1}}{\sqrt{f_{i1}}}-\frac{\gamma_{i2}}{\sqrt{f_{i2}}}\right]+\varepsilon^{\star\star}_{i1k}, \\ Y^{\star\star}_{i2k} &= \frac{1}{\sqrt{2(\sigma^2+\sigma_{i12})}}\left[(\frac{1}{\sqrt{f_{i1}}}+\frac{1}{\sqrt{f_{i2}}})\mu+(\frac{1}{\sqrt{f_{i1}}}+\frac{1}{\sqrt{f_{i2}}})\alpha_i\right. \\ &\quad \left.+\frac{\beta_1}{\sqrt{f_{i1}}}+\frac{\beta_2}{\sqrt{f_{i2}}}+\frac{\gamma_{i1}}{\sqrt{f_{i1}}}+\frac{\gamma_{i2}}{\sqrt{f_{i2}}}\right]+\varepsilon^{\star\star}_{i2k} \end{aligned}$$

where ε's are uncorrelated and with variance 1.

This model is a typical regression model and easily solved by normal equations, that is, by simple matrix operations to obtain estimators for μ, α_i, β_j and γ_{ij}.

Now we are ready to discuss the model we are interested.

$$Y_{ijk} = \mu+\alpha_i+\beta_j+\gamma_{ij}+\varepsilon_{ijk}$$

for $i=1,\cdots,a$; $j=1,\cdots,b$; and $k=1,\cdots,n$.

$$\mathrm{Cov}(\varepsilon_{i_1,j_1,k_1},\varepsilon_{i_2,j_2,k_2}) = \begin{cases} 0, & \text{if } i_1\neq i_2 \\ 0, & \text{if } k_1\neq k_2 \\ \sigma^2 f_{i_1,j_1}, & \text{if } i_1=i_2, j_1=j_2, \text{ and } k_1=k_2 \\ \sigma_{i12}\sqrt{f_{i_1,j_1}f_{i_2,j_2}}, & \text{if } i_1=i_2, j_1\neq j_2, \text{ and } k_1=k_2 \end{cases}$$

where

$$f_{i,j} \quad = \quad \mu + \alpha_i + \beta_j + \gamma_{ij},$$

that is, the variance is proportional to the mean. Obviously, there is no simple way to find the maximum likelihood estimator or other good estimators.

5.2.2 Numerical Steps for Finding the Estimators of Parameters

We propose the following numerical steps to find the estimators.

Step 1: Initial estimators:

Assume that $f_{ij} = 1$ and find the estimators.

$$\begin{aligned} Y^{\star\star}_{i1k} &= \frac{1}{\sqrt{2(\sigma^2 - \sigma_{i12})}} \left[\beta_1 - \beta_2 + \gamma_{i1} - \gamma_{i2}\right] + \varepsilon^{\star\star}_{i1k}, \\ Y^{\star\star}_{i2k} &= \frac{1}{\sqrt{2(\sigma^2 + \sigma_{i12})}} \left[\mu/2 + \alpha_i/2 + \beta_1 + \beta_2 + \gamma_{i1} + \gamma_{i2}\right] + \varepsilon^{\star\star}_{i2k}. \end{aligned}$$

Step 2:

Based on the estimators of μ, α_i, β_j and γ_{ij}, we obtain the mean estimator for $\mu + \alpha_i + \beta_j + \gamma_{ij}$. Substituting these estimators to calculate f_{ij}. Assume that f_{ij} were constant (of course they are not) and obtain the estimators:

$$\begin{aligned} Y^{\star\star}_{i1k} = \frac{1}{\sqrt{2(\sigma^2 - \sigma_{i12})}} \Bigg[& \left(\frac{1}{\sqrt{\hat{\mu} + \hat{\alpha}_i + \hat{\beta}_1 + \hat{\gamma_{i1}}}} - \frac{1}{\sqrt{\hat{\mu} + \hat{\alpha}_i + \hat{\beta}_2 + \hat{\gamma_{i1}}}} \right) \mu \\ & + \left(\frac{1}{\sqrt{\hat{\mu} + \hat{\alpha}_i + \hat{\beta}_1 + \hat{\gamma_{i1}}}} - \frac{1}{\sqrt{\hat{\mu} + \hat{\alpha}_i + \hat{\beta}_2 + \hat{\gamma_{i2}}}} \right) \alpha_i \\ & + \frac{\beta_1}{\sqrt{\hat{\mu} + \hat{\alpha}_i + \hat{\beta}_1 + \hat{\gamma_{i1}}}} - \frac{\beta_2}{\sqrt{\hat{\mu} + \hat{\alpha}_i + \hat{\beta}_2 + \hat{\gamma_{i2}}}} \\ & + \frac{\gamma_{i1}}{\sqrt{\hat{\mu} + \hat{\alpha}_i + \hat{\beta}_1 + \hat{\gamma_{i1}}}} - \frac{\gamma_{i2}}{\sqrt{\hat{\mu} + \hat{\alpha}_i + \hat{\beta}_2 + \hat{\gamma_{i2}}}} \Bigg] + \varepsilon^{\star\star}_{i1k} \end{aligned}$$

$$
\begin{aligned}
Y_{i2k}^{\star\star} \;=\; & \frac{1}{\sqrt{2(\sigma^2+\sigma_{i12})}}\left[\left(\frac{1}{\sqrt{\hat{\mu}+\hat{\alpha}_i+\hat{\beta}_1+\hat{\gamma_{i1}}}}+\frac{1}{\sqrt{\hat{\mu}+\hat{\alpha}_i+\hat{\beta}_2+\hat{\gamma_{i1}}}}\right)\mu\right. \\
& +\left(\frac{1}{\sqrt{\hat{\mu}+\hat{\alpha}_i+\hat{\beta}_1+\hat{\gamma_{i1}}}}+\frac{1}{\sqrt{\hat{\mu}+\hat{\alpha}_i+\hat{\beta}_2+\hat{\gamma_{i2}}}}\right)\alpha_i \\
& +\frac{\beta_1}{\sqrt{\hat{\mu}+\hat{\alpha}_i+\hat{\beta}_1+\hat{\gamma_{i1}}}}+\frac{\beta_2}{\sqrt{\hat{\mu}+\hat{\alpha}_i+\hat{\beta}_2+\hat{\gamma_{i2}}}} \\
& \left.+\frac{\gamma_{i1}}{\sqrt{\hat{\mu}+\hat{\alpha}_i+\hat{\beta}_1+\hat{\gamma_{i1}}}}+\frac{\gamma_{i2}}{\sqrt{\hat{\mu}+\hat{\alpha}_i+\hat{\beta}_2+\hat{\gamma_{i2}}}}\right]+\varepsilon_{i2k}^{\star\star}
\end{aligned}
$$

Step 3: Execute Step 2 again until the estimators converge.

When finished, the estimators are those we are interested.

Chapter 6

Patterns of Nucleotide Substitution Rates at Multiple Loci of Animal Mitochondrial Genomes

6.1 Introduction

Animal cells hold two separate genomes located in the nucleus and mitochondria. Nuclear genome stores the majority of genetic information, while mitochondria stores an additional portion. Mitochondrion is a small two-membrane-bound organelle that is the source of cell's power. The animal mitochondrial genomes are normally circular, $15-20$ KB in length, and encode 13 proteins used for energy production. The two genomes have distinct differences in heritage and evolutionary processes. First, Brown *et al.* (1997) showed that substitution rates were much lower in nuclear genome than in its counterpart of mitochondria. Mitochondria use a different genetic code from nucleus. Most mitochondrial mRNAs cannot be translated in the nuclear-cytoplasmic compartment. Hence, this specific mtDNA genetic code confines the mtDNA genes to express within the mitochondria. As well known, each parent gives their offspring approximate half of their nuclear DNA. In contrast, offspring only receive their mothers' mitochondrial DNA, but none of their father's mtDNA, with no recombination. The result is that mtDNA is passed on only along the maternal line.

Mitochondrial DNA was believed to be an ideal molecular "clock". Two reasons mainly contributed to this belief. The first reason is that mtDNA is not divided during cell division and generally passed down via the mother's line. Secondly, most mutation in mtDNA was thought to be neutral instead of natural selection. Therefore mtDNA was widely used to

date the origin of taxonomic group through the use of molecular clocks. New evidences have indicated that mtDNA is subject to natural selection (Fos *et al.* 1990, Malhotra *et al.* 1994). Recent studies also showed that paternal mtDNA can on rare occasions enter an egg during fertilization and alter the maternal mtDNA through recombination (Ladoukakis *et al.* 2001, Meunier *et al.* 2001). Such recombination would affect the mutation rate of mtDNA. Those evidences as well as the fact of fast mutation rate in mtDNA suggest that mitochondrial DNA probably is not an idea molecular "clock". It is doubt to date the origin by only counting the number of mutants in mtDNA between any taxonomic groups.

Many studies have been conducted to examine the patterns of nucleotide substitution at animal mitochondrial loci (*e.g.*, Hasegawa *et al.* 1989, Fos *et al.* 1990, Malhotra *et al.* 1994, Adachi *et al.* 1995, Parsons *et al.* 1997) by using single-locus approach. Here we use muti-loci approach to simultaneously study the patterns of nucleotide substitution at multiple loci of animal mitochondrial genomes. As we introduced previously, variation in substitution rates can be governed by three components: locus effect, lineage effect and interplay between these two effects. In general, locus effect and lineage effect have been well documented (*e.g.,* MacRea *et al.* 1988, Malhotra *et al.* 1994), but the interaction between them have not been widely studied, especially for nucleotide substitution rates of animal mitochondrial sequences.

In this study, we examined patterns in substitution rates of the 11 protein coding genes from five mtDNA data sets. Each data set includes three taxa, and together the five data sets represent a broad range of animal species. Pairwise comparisons were conducted among the 11 genes for each data set separately to identify the factors that might affect nucleotide substitution rates in animal mitochondrial genes. There are two purposes of the study. First, we want to know whether locus effects, lineage effects and locus $\times$ lineage effects can be detected in nucleotide substitution rates of animal mtDNA. Secondly, we want to know what kinds of molecular evolutionary processes may affect animal mtDNA.

6.2 Materials and Methods

6.2.1 The mtDNA Sequences

The data are from Hilton *et al.* (1996). The mtDNA sequence data they used satisfied the following criteria. First, the whole mtDNA sequences must have been identified from three-taxon phylogenies. To insure that each data set can represent the independent process of evolution, they tried to reduce phylogenetic overlap in the three-taxon data sets.

We studied patterns of nucleotide substitution rates in animal mtDNA from five data sets. Table 6.1 provides their classifications, scientifical and common names, accession numbers in GenBank and references. The first data set contains two strains of mosquitoes susceptible to malaria from the family Diptera, and with fruit fly as an outgroup. The second data set contains two types of fish from the superfamily Cyprinoidea, with trout as an outgroup (Weber *et al.* 1916). The third data set contains two birds in the ratite family, with chicken as an outgroup. The fourth data set contains two primates, human and gorilla, with orangutan as an outgroup (Kavanagh 1984, Hasegawa *et al.* 1985). The last data set contains two mammals, cat and seal, in the Order Carnivora, with rhino from the Order Perissondactyla as an outgroup (Stains 1984).

The 11 genes of the 13 protein coding genes were examined in five data sets. The gene *atp*8 was not included in analysis because it is too short ($160 - 200bp$), and the gene *nd*6 was not used because it was difficult to align. The DNA sequences were aligned manually, and overlay between genes was eliminated from analysis. Table 6.2 gives the gene names and lengths in comparisons among five organisms. The lengths of the 11 genes range from $201bp$ at the gene *nd*3 for Bird to $1836bp$ at the gene *nd*5 for Fish. In general, gene *nd*5 has the longest length, while gene *nd*4*l* has the shortest length. Similar lengths present at the same gene across the different organisms except for the genes *nd*4*l* and *nd*5 for Insect and the gene *nd*3 for Bird.

Table 6.1: List of species whose mtDNA sequences were studies

Organism	Scientific name	Common name	Acce. No.	Reference
Insect	Anopheles quadrimaculatus A	mosquito	L04272	Cockburn 1990
	Anopheles gambiae	mosquito	L20934	Beard 1993
	Drosophila melanogaster*	fruit fly	U37541	Lewis 1995
Fish	Cyrinus carpio	carp	X61010	Chang 1994
	Crossostoma lacustre	loach	M91245	Tzeng 1992
	Oncorhynchus mykiss*	trout	L29771	Zardoya 1995
Bird	Rhea americana	Rhea	Y16884	Harlid 1998
	Struthio camelus	ostrich	Y12025	Harlid 1997
	Gallus gallus*	chicken	NC001323	Desjardins 1990
Primate	Homo sapian	human	X93334	Arnason 1996
	Garilla gorilla	gorilla	X93347	Xu 1996a
	Pongo pygmaeus*	orangutan	D38115	Horai 1992
Mammal	Felis catus	cat	U20753	Lopez 1996
	Halichoerus grypus	seal	X72004	Arnason 1993
	Rhinoceros unicornis*	rhino	X97336	Xu 1996b

Star (*) represents an outgroup.

Table 6.2: Sequence lengths (*bp*) at 11 loci for different organisms

Gene	Insect	Fish	Bird	Primate	Mammal
atp6	678	675	684	633	636
co1	1527	1551	1548	1539	1539
co2	678	690	681	681	684
co3	789	786	783	783	783
cytb	1124	1137	1137	1140	1140
nd1	933	975	975	950	954
nd2	1026	1057	1038	1041	1041
nd3	354	336	201	345	345
nd4	1335	1371	1371	1377	1377
nd4l	243	291	297	294	297
nd5	1725	1836	1816	1806	1806

6.2.2 Statistical Methods

The purpose of the study was to detect the factors that influence nucleotide substitution rates in animal mitochondrial genes. To accomplish our goal, we employed proposed ANOVA-based method, ANOVA.L.C, to the mtDNA data. We used the HKY85 Model to estimate the branch lengths since rates of transition and transversion differ considerably in mtDNA sequences (Hasegawa *et al.* 1985). Previous simulation studies have suggested that we would use small number of replicates to obtain high statistical powers for ANOVA.L.C, thus we separated each sequence into two data sets in an alternating way. Furthermore, since each codon is a single replication of the evolution experiment, we split the sequences of genes into two data sets representing alternated codons. There are some gaps presented in the mtDNA data sets, we therefore cleaned those gaps before separating the sequences. To guarantee the codon at the same position to be assigned into the same separated data set, we deleted the three nucleotides of a codon in the three species at once as long as there was any gap presented in that codon.

6.3 Results and Discussion

We applied ANOVA.L.C-based relative ratio test to all 275 pairs of genes within each of the five data set (Table 6.1). The null hypothesis of the relative ratio test is that locus and lineage effects are sufficient to explain the data; rejection of the null hypothesis suggests that the presence of locus $\times$ lineage effects. As mentioned in Chapter 2, one of merit of ANOVA-based relative ratio test is that locus and lineage effects can be easily detected in the meantime. To better understand the factors that influence substitution rates of mitochondrial genes, we reported the results of locus $\times$ lineage effects as well as locus and lineage effects for each pairwise comparison. The results are summarized in Table 6.3 $-$ 6.7. In each table, stars indicate significant tests at $p< 0.05$(*) or $p< 0.01$(**) levels. The significant tests for locus $\times$ lineage effects are above the diagonal, while the significant tests for locus and lineage (inside the parentheses) effects are below the diagonal.

The total number 22 out of 275 (8%) relative ratio tests are significant at the 0.05

level. This gives the impression that locus × lineage effects are not common in animal mitochondrial genomes. For example, none of relative ratio comparisons within bird data set is significant at the 0.05 level. The fish data set has the highest frequency of significant relative ratio results at 10 out of 55 (18%). Secondly, a common pattern of significant relative ratio tests does not present among five data sets. For example, comparisons involving *nd4l* gene reject relative ratio tests most often in the primate data set and do not reject in the others, suggesting that this loci may have atypical rate in primate relative to other loci. There are also many significant rejections for both locus and lineage effects for comparisons involving *nd4l* in the primate data. Together we may first need to check whether or not the sequence alignment for this locus is fine.

Insect Two strains of mosquitoes susceptible to malaria are under investigation. Only are two comparisons involving *nd*5 gene detected to reject relative ratio tests at the significance level 0.05. This indicates that rates of nucleotide substitutions for the two mosquitoes are well conserved between *nd*5 with both *cytb* and *nd*4 genes. For testing locus and lineage effects, a total of 4 comparisons involving *co*1 gene gives significant results for locus effects. Two significant lineage effects are found between *nd*3 and *atp*6, and between *co*1 and *nd*4. There is fairly convincing evidence that the two mosquitoes don't behave very differently.

Fish Comparisons between fish carp and loach reject the null hypothesis for 18% of relative ratio tests which include 3 tests involving *nd*2, 3 tests involving *nd*5, 2 tests involving *nd*4, and 1 between *nd*2 and *nd*3. A large number of locus and lineage effects are detected to be significant. For example, comparisons involving *atp*6 gene reject the null hypothesis of no lineage effect at 60%, and reject the null hypothesis of no locus effect at 40%. Comparisons involving *co*3 gene have the highest frequency of significant both locus and lineage effects at 30%. Together results among relative ratio tests, tests of locus effects, and tests of lineage effects, provide compelling evidence that carp and loach have experienced different evolutionary forces since they diverged from the latest common ancestors.

Bird There is no rejection of relative ratio tests for Rhea and ostrich across the 11 genes at the significance level 0.05. This suggests that rates of substitution for both birds are well conserved across the 11 genes. Comparisons involving *cytb* with *nd*1, *nd*2, and *nd*4 reject the null hypothesis of no locus effect at significance level 0.01, and only is one rejection detected between *cytb* and *nd*2. Comparisons involving *co*2 with *nd*1, *nd*2, and *nd*4 only reject the null hypothesis of no locus effect at the significance level 0.05. Locus effects are found to be significant for the tests involving *co*1 with *nd*1 and *nd*2, while lineage effect is significant between Rhea and ostrich for the comparison between *co*1 and *nd*4. This provides evidence that locus effects are predominate between a set of three genes *co*1, *co*2 and *cytb* with a set of three genes *nd*1, *nd*2 and *nd*4.

Primate Comparisons between human and gorilla detect 50% rejections of relative ratio tests and 60% rejections of the null hypothesis of no locus effects, and 40% rejections of the null hypothesis of no lineage effect for tests involving *nd*4*l* gene. It is clear that *nd*4*l* gene does have quite different evolutionary behavior. Two significant rejections of relative ratio tests are also detected for comparisons involving *nd*5 gene, and comparisons involving *nd*3 genes.

Mammal Comparisons between cat and seal detect a total of one rejection of relative ratio tests between *nd*5 and *cytb* at the significance level 0.05. Many locus and lineage effects are significant at either level 0.05 or 0.01. For example, comparisons involving *nd*1 reject the null hypothesis of no locus effect for 50%, while comparisons involving *nd*5 reject the null hypothesis of no locus effect for 40% at the significance level 0.05. For lineage effects, comparisons involving *cytb* reject the null hypothesis of no lineage effect for 40%, while comparisons involving *cytb* and *nd*4 reject the null hypothesis of no lineage effect for 30% at the significance level 0.05. This suggests that locus and lineage effects play an important role for the mitochondrial genes since cat and seal diverged from their latest ancestor.

Discussion We have examined nucleotide substitution rates in animal mitochondrial genes using ANOVA.L.C method. The locus effects, lineage effects and locus $\times$ lineage effects have been detected in animal mtDNA. The results provide evidences that lineage and locus effects are more common components than locus $\times$ lineage effects in variations of nucleotide substitution rates of animal mitochondrial genomes. Furthermore, the results also suggest that these effects vary qualitatively among phylogenetically distinct data sets. To better understand the evolutionary forces affecting animal mtDNA, the functions of mitochondrial genes will be first discussed.

Mitochondria supply energy to the cell (ATP) produced by the respiratory chain by using oxygen. The respiratory chain consists of the four succeeding complexes: complex I, complex II, complex III and complex IV. All the mitochondrial DNA genes encode subunits of the oxidative phosphorylation enzymes, except for the complex II. Complex I consists of approximately 42 polypeptides, seven ($nd1, nd2, nd3, nd4, nd4l, nd5, nd6$) encoded by the mtDNA; Complex III of about 11 polypeptides, one ($cytb$) encoded by mtDNA; Complex IV of 13 polypeptides, three ($co1, co2, co3$) encoded by the mtDNA; Complex V of 12 polypeptides, two ($atp6, atp8$) encoded by the mtDNA. However, while mtDNA does not code for any DNA repair proteins, it has been observed that a number of repair factors can be found in mitochondrial extracts (Bohr *et al.* 1999). Thus, this could indicate the presence of a more complex repair process in mtDNA than in nuclear DNA. A variety of studies in apoptosis focusing on mitochondria suggested that mitochondria might have a pivotal part in controlling cell life and death (*e.g.,* Green *et al.* 1998, Murphy *et al.* 1999, Rustin *et al.* 2000). Those functions of mitochondrial genes indicate that the metabolic rate of an organism and the process of aging might relate to functions of mitochondria. It is known that metabolism is related to the size of a creature and whether it is warm-blooded or cold-blooded (Martin *et al.* 1993). Among warm-blooded animals, bigger animals tend to have a slower metabolism than smaller ones. Cold-blooded animals have a slower metabolism than warm-blooded animals.

In our survey of the five organisms, four are warm-blooded animals except for fish.

Among the four warm-blooded animals, there exist big differences in weight within bird, primate and mammal. In bird, ostrich is the largest living bird in the world (average $90-130kg$) and rhea is much smaller than ostrich. Ostrich has relative longer life span than rhea. In primate, gorilla is larger than human in general. In mammal, seal is much larger than cat. As discussed in Chapter 1, different metabolic rates can result in lineage effects. Those lineage effects detected in bird, Primate and mammal probably can be explained by different metabolisms. Currently, there are two kinds of interpretations for lineage effects. The first one assumes that lineage effects are caused by differences in mutation rates between evolutionary lineages. That is, the variation in mutation rates could reflect differences in metabolic process (Martin *et al.* 1992, Martin *et al.* 1993). The second one uses different selective pressure between evolution lineages to explain lineage effects. For example, adaptation to different temperatures like cold-blooded vs. warm-blooded animals or differences in population size. The differences in substitution rates between evolution lineages have been found for cold-blooded vs. warm-blooded animal (Martin *et al.*1992).

There is a relatively large number of locus effects that have been detected in our five data. Locus effects might be caused by the differences in mutation rates for a particular locus. Generally, the mutation rates for animal mtDNA can be divided by three catalogues: the rapidly evolving sequences ($nd1, nd3, nd4, nd5$, and $nd6$), the moderately evolving sequences ($cytb$), and the slowly evolving sequences ($co1, co2, co3, atp6, nd2$ and $nd4l$) (Billington, 2002). Currently, the evolutionary process contributing to locus $\times$ lineage effects are not very clear. One possible explanation is the levels of selective pressures vary among genes and among evolutionary lineages. Note the gene $nd4l$ has very interesting results for primate data: comparisons involving that gene reject relative ratio test most often in the primate data and do not reject in other data as well as for locus and lineage effects. The gene $nd4l$ is one of the seven nd genes involving complex I in respiratory chain. A number of studies have been shown that any disorder in complex I could result in dysfunction in brain (*e.g.,* Wiedemann *et al.* 2002, Mezies *et al.* 2002, Chalmers 2002). The fact that human has higher intelligence than gorilla may result in the atypical results for primate data. In other

words, *nd4l* is probably more likely to involve the activities of brains than the other genes.

Table 6.3: Results of ANOVA.L.C-based relative ratio test for insect

Locus	atp6	co1	co2	co3	cytb	nd1	nd2	nd3	nd4	nd4l	nd5
atp6	–										
co1		–									
co2			–								
co3		*		–							
cytb		*			–						*
nd1		*				–					
nd2							–				
nd3	(*)							–			
nd4		*							–		*
nd4l										–	
nd5		(*)									–

Note: interaction effects above diagonal, locus and lineage (inside parentheses) below diagnal. Stars indicate significant tests at the $p < 0.05$(*) or $p < 0.01$(**).

Table 6.4: Results of ANOVA.L.C-based relative ratio test for fish

Locus	atp6	co1	co2	co3	cytb	nd1	nd2	nd3	nd4	nd4l	nd5
atp6	–	*					*		*		*
co1	*	–									
co2			–								
co3	*(*)			–			*		*		*
cytb	(*)				–						
nd1	(*)			*(*)		–	*				
nd2	*	**		**	*		–	*			
nd3	(*)			*(*)		(*)		–			*
nd4		*			*				–		
nd4l							*		*	–	
nd5	*	**(*)		**	**		(*)			*	–

Table 6.5: Results of ANOVA.L.C-based relative ratio test for bird

Locus	atp6	co1	co2	co3	cytb	nd1	nd2	nd3	nd4	nd4l	nd5
atp6	–										
co1		–									
co2			–								
co3				–							
cytb					–						
nd1		*	*		**	–					
nd2		*	**		**(*)		–				
nd3								–			
nd4		(*)	*		**				–		
nd4l										–	
nd5											–

Table 6.6: Results of ANOVA.L.C-based relative ratio test for primate

Locus	atp6	co1	co2	co3	cytb	nd1	nd2	nd3	nd4	nd4l	nd5
atp6	–							*		**	
co1		–								*	
co2			–								
co3				–							
cytb					–					*	
nd1					*	–		*		**	
nd2							–				
nd3					(*)			–			*
nd4									–	*	
nd4l	**(*)	*	(*)		**(*)	*	*	**(**)	*	–	**
nd5		*				*				**	–

Table 6.7: Results of ANOVA.L.C-based relative ratio test for mammal

Locus	atp6	co1	co2	co3	cytb	nd1	nd2	nd3	nd4	nd4l	nd5
atp6	–										
co1	*	–									
co2			–								
co3	*			–							
cytb	(*)				–						*
nd1	**(*)				*(*)	–					
nd2				*		**	–				
nd3						*		–			
nd4	(**)	*		*	(*)	**(*)			–		
nd4l						*				–	
nd5		*	*	**	(*)	**					–

Appendix A

Deriving Expected Mean Squares for the two-way ANOVA Model with Covariance Structure

Consider the model

$$Y_{ijk} = \mu + \alpha_i + \beta_j + \gamma_{ij} + \varepsilon_{ijk}$$

for $i = 1, \cdots, a$; $j = 1, \cdots, b$; and $k = 1, \cdots, n$.

Assume that

$$\mathrm{Cov}(\varepsilon_{i_1,j_1,k_1}, \varepsilon_{i_2,j_2,k_2}) = \begin{cases} 0, & \text{if } i_1 \neq i_2; \\ 0, & \text{if } k_1 \neq k_2; \\ \sigma^2, & \text{if } i_1 = i_2,\ j_1 = j_2, \text{ and } k_1 = k_2; \\ \sigma_{i12} < 0, & \text{if } i_1 = i_2 = i,\ j_1 \neq j_2, \text{ and } k_1 = k_2. \end{cases}$$

and $E[\varepsilon_{ijk}] = 0$, with the following restrictions:

$$\sum_{i=1}^{a} \alpha_i = 0, \quad \sum_{j=1}^{b} \beta_j = 0, \quad \sum_{i=1}^{a} \gamma_{ij} = 0, \quad \text{and} \quad \sum_{j=1}^{b} \gamma_{ij} = 0. \tag{A.1}$$

Note that

$$\begin{aligned}
\overline{Y}_{...} &= \mu + \overline{\varepsilon}_{...} \\
\overline{Y}_{i..} &= \mu + \alpha_i + \overline{\varepsilon}_{i..} \\
\overline{Y}_{.j.} &= \mu + \beta_j + \overline{\varepsilon}_{.j.} \\
\overline{Y}_{ij.} &= \mu + \alpha_i + \beta_j + \gamma_{ij} + \overline{\varepsilon}_{ij.}
\end{aligned}$$

and

$$\begin{aligned}
SSA &= nb \sum_{i=1}^{a} \left(\overline{Y}_{i..} - \overline{Y}_{...}\right)^2 \\
SSB &= na \sum_{j=1}^{b} \left(\overline{Y}_{.j.} - \overline{Y}_{...}\right)^2 \\
SSAB &= n \sum_{i=1}^{a} \sum_{j=1}^{b} \left(\overline{Y}_{ij.} - \overline{Y}_{i..} - \overline{Y}_{.j.} + \overline{Y}_{...}\right)^2 \\
SSE &= \sum_{i=1}^{a} \sum_{j=1}^{b} \sum_{k=1}^{n} \left(Y_{ijk} - \overline{Y}_{ij.}\right)^2
\end{aligned}$$

$$\begin{aligned}
\frac{1}{nb} SSA &= \sum_{i=1}^{a} \left[\alpha_i + (\overline{\varepsilon}_{i..} - \overline{\varepsilon}_{...})\right]^2 \\
&= \sum_{i=1}^{a} \alpha_i^2 + \sum_{i=1}^{a} (\overline{\varepsilon}_{i..} - \overline{\varepsilon}_{...})^2 + \sum_{i,j} \alpha_i (\overline{\varepsilon}_{j..} - \overline{\varepsilon}_{...})
\end{aligned}$$

$$E\left[\frac{1}{nb} SSA\right] = \sum_{i=1}^{a} \alpha_i^2 + E\left[\sum_{i=1}^{a} (\overline{\varepsilon}_{i..} - \overline{\varepsilon}_{...})^2\right]$$

$$\begin{aligned}
E\left[(\overline{\varepsilon}_{i..} - \overline{\varepsilon}_{...})^2\right] &= \mathrm{Var}\left(\overline{\varepsilon}_{i..} - \overline{\varepsilon}_{...}\right) \\
&= \mathrm{Var}\left(\overline{\varepsilon}_{i..}\right) + \mathrm{Var}\left(\overline{\varepsilon}_{...}\right) - 2 \cdot \mathrm{Cov}\left(\overline{\varepsilon}_{i..}, \overline{\varepsilon}_{...}\right)
\end{aligned}$$

$$
\begin{aligned}
\mathrm{Var}\left(\overline{\varepsilon}_{i..}\right) &= \mathrm{Var}\left[\sum_{j,k} \varepsilon_{ijk}/(bn)\right] \\
&= \frac{1}{b^2n^2}\mathrm{Var}\left(\sum_{j,k} \varepsilon_{ijk}\right) \\
&= \frac{1}{b^2n^2}\left[\sum_{j,k} \mathrm{Var}\left(\varepsilon_{ijk}\right) + \sum_{(j_1,k_1)\neq(j_2,k_2)} \mathrm{Cov}(\varepsilon_{i,j_1,k_1}, \varepsilon_{i,j_2,k_2})\right] \\
&= \frac{1}{b^2n^2}\left[bn\sigma^2 + \sum_{j_1\neq j_2} \mathrm{Cov}(\varepsilon_{i,j_1,k}, \varepsilon_{i,j_2,k})\right] \\
&= \frac{1}{b^2n^2}\left[bn\sigma^2 + n\cdot\sum_{j_1\neq j_2} \sigma_{i12}\right] \\
&= \frac{1}{b^2n^2}\left[bn\sigma^2 + n(b^2-b)\sigma_{i12}\right] \\
&= \frac{1}{bn}\left[\sigma^2 + (b-1)\sigma_{i12}\right]
\end{aligned}
$$

$$
\begin{aligned}
\mathrm{Var}(\overline{\varepsilon}_{...}) &= \mathrm{Var}\left(\frac{1}{a}\sum_i \overline{\varepsilon}_{i..}\right) \\
&= \frac{1}{a^2}\mathrm{Var}\left(\sum_i \overline{\varepsilon}_{i..}\right) \\
&= \frac{1}{a^2}\sum_i \mathrm{Var}\left(\overline{\varepsilon}_{i..}\right) \\
&= \frac{1}{a^2bn}\sum_i \left[\sigma^2 + (b-1)\sigma_{i12}\right] \\
&= \frac{1}{a^2bn}\left[a\sigma^2 + (b-1)\sum_i \sigma_{i12}\right] \\
&= \frac{1}{abn}\left[\sigma^2 + \frac{b-1}{a}\sum_i \sigma_{i12}\right]
\end{aligned}
$$

$$\begin{aligned}\operatorname{Cov}\left(\overline{\varepsilon}_{i..}, \overline{\varepsilon}_{...}\right) &= \operatorname{Cov}\left(\overline{\varepsilon}_{i..}, \frac{1}{a}\sum_{t=1}^{a}\overline{\varepsilon}_{t..}\right) \\ &= \frac{1}{a}\operatorname{Var}(\overline{\varepsilon}_{i..})\end{aligned}$$

$$\begin{aligned}E\left(\overline{\varepsilon}_{i..} - \overline{\varepsilon}_{...}\right)^2 &= \operatorname{Var}\left(\overline{\varepsilon}_{i..}\right) + \operatorname{Var}\left(\overline{\varepsilon}_{...}\right) - 2\cdot\operatorname{Cov}\left(\overline{\varepsilon}_{i..}, \overline{\varepsilon}_{...}\right) \\ &= \frac{a-2}{a}\operatorname{Var}\left(\overline{\varepsilon}_{i..}\right) + \operatorname{Var}\left(\overline{\varepsilon}_{...}\right)\end{aligned}$$

$$\begin{aligned}\sum_i E\left(\overline{\varepsilon}_{i..} - \overline{\varepsilon}_{...}\right)^2 &= \frac{a-2}{a}\sum_i \operatorname{Var}\left(\overline{\varepsilon}_{i..}\right) + a\operatorname{Var}\left(\overline{\varepsilon}_{...}\right) \\ &= \frac{a-2}{abn}\left[a\sigma^2 + (b-1)\sum_i \sigma_{i12}\right] + \frac{a}{abn}\left[\sigma^2 + \frac{b-1}{a}\sum_i \sigma_{i12}\right] \\ &= \frac{a-1}{bn}\left[\sigma^2 + \frac{b-1}{a}\sum_i \sigma_{i12}\right]\end{aligned}$$

Therefore,

$$E\left[\frac{1}{nb}SSA\right] = \sum_{i=1}^{a}\alpha_i^2 + \frac{a-1}{bn}\left[\sigma^2 + \frac{b-1}{a}\sum_i \sigma_{i12}\right]$$

and

$$E\left[MSA\right] = \frac{nb}{a-1}\sum_i^a \alpha_i^2 + \sigma^2 + \frac{b-1}{a}\sum_i^a \sigma_{i12}$$

Now, find $E[MSB]$

$$\begin{aligned}\frac{1}{na}SSB &= \sum_{j=1}^{b}\left[\beta_j + (\overline{\varepsilon}_{.j.} - \overline{\varepsilon}_{...})\right]^2 \\ &= \sum_{j=1}^{b}\beta_j^2 + \sum_{j=1}^{b}(\overline{\varepsilon}_{.j.} - \overline{\varepsilon}_{...})^2 + \sum_{j}\beta_j(\overline{\varepsilon}_{.j.} - \overline{\varepsilon}_{...})\end{aligned}$$

$$E\left[\frac{1}{na}SSB\right] = \sum_{j=1}^{b}\beta_j^2 + E\left[\sum_{j=1}^{b}(\overline{\varepsilon}_{.j.} - \overline{\varepsilon}_{...})^2\right]$$

$$\begin{aligned}E\left[(\overline{\varepsilon}_{.j.} - \overline{\varepsilon}_{...})^2\right] &= \mathrm{Var}\left(\overline{\varepsilon}_{.j.} - \overline{\varepsilon}_{...}\right) \\ &= \mathrm{Var}\left(\overline{\varepsilon}_{.j.}\right) + \mathrm{Var}\left(\overline{\varepsilon}_{...}\right) - 2\mathrm{Cov}\left(\overline{\varepsilon}_{.j.}, \overline{\varepsilon}_{...}\right)\end{aligned}$$

$$\mathrm{Var}\left(\overline{\varepsilon}_{.j.}\right) = \mathrm{Var}\left[\sum_{i,k}\varepsilon_{ijk}/na\right] = \frac{1}{na}\sigma^2$$

$$
\begin{aligned}
\operatorname{Cov}\left(\overline{\varepsilon}_{.j.}, \overline{\varepsilon}_{...}\right) &= \operatorname{Cov}\left(\overline{\varepsilon}_{.j.}, \frac{1}{b}\sum_{t=1}^{b}\overline{\varepsilon}_{.t.}\right) \\
&= \frac{1}{b}\left[\operatorname{Var}(\overline{\varepsilon}_{.j.}) + \sum_{t\neq j}\operatorname{Cov}(\overline{\varepsilon}_{.j.}, \overline{\varepsilon}_{.t.})\right] \\
&= \frac{1}{abn}\sigma^2 + \frac{1}{a^2bn^2}\sum_{t\neq j}\left[\operatorname{Cov}(\sum_{i_1,k_1}\varepsilon_{i_1jk_1}, \sum_{i_2,k_2}\varepsilon_{i_2tk_2})\right] \\
&= \frac{1}{abn}\sigma^2 + \frac{1}{a^2bn^2}\sum_{t\neq j}\left[\sum_{\substack{i_1,k_1\\ i_2,k_2}}\operatorname{Cov}(\varepsilon_{i_1jk_1}, \varepsilon_{i_2jk_2})\right] \\
&= \frac{1}{abn}\sigma^2 + \frac{1}{a^2bn^2}\sum_{t\neq j}\left[\sum_{i,k}\operatorname{Cov}(\varepsilon_{ijk}, \varepsilon_{itk})\right] \\
&= \frac{1}{abn}\sigma^2 + \frac{1}{a^2bn^2}\sum_{t\neq j}\left(n\sum_{i=1}^{a}\sigma_{i12}\right) \\
&= \frac{1}{abn}\sigma^2 + \frac{1}{a^2bn^2}(b-1)n\sum_{i=1}^{a}\sigma_{i12} \\
&= \frac{1}{abn}\left(\sigma^2 + \frac{b-1}{a}\sum_{i=1}^{a}\sigma_{i12}\right)
\end{aligned}
$$

$$
\begin{aligned}
E\left[(\overline{\varepsilon}_{.j.} - \overline{\varepsilon}_{...})^2\right] &= \operatorname{Var}(\overline{\varepsilon}_{.j.}) + \operatorname{Var}\left(\overline{\varepsilon}_{...}\right) - 2\cdot\operatorname{Cov}\left(\overline{\varepsilon}_{.j.}, \overline{\varepsilon}_{...}\right) \\
&= \operatorname{Var}(\overline{\varepsilon}_{.j.}) - \operatorname{Var}(\overline{\varepsilon}_{...})
\end{aligned}
$$

$$\sum_{j=1}^{b} E\left[(\overline{\varepsilon}_{.j.} - \overline{\varepsilon}_{...})^2\right] = \sum_{j}\left[\frac{1}{na}\sigma^2 - \frac{1}{abn}\left(\sigma^2 + \frac{b-1}{a}\sum_{i=1}^{a}\sigma_{i12}\right)\right]$$
$$= \frac{b-1}{na}\left(\sigma^2 - \frac{1}{a}\sum_{1}\sigma_{i12}\right)$$

Therefore,

$$E\left[\frac{1}{na}SSB\right] = \sum_{j=1}^{b}\beta_j^2 + \frac{b-1}{na}\left(\sigma^2 - \frac{1}{a}\sum_{1}\sigma_{i12}\right)$$

and

$$E\left[MSB\right] = \frac{na}{b-1}\sum_{j}\beta_j^2 + \sigma^2 - \frac{1}{a}\sum_{i}\sigma_{i12}$$

Then we need to find $E\left[MSAB\right]$

$$\frac{1}{n}SSAB = \sum_{i=1}^{a}\sum_{j=1}^{b}(\gamma_{ij} + \overline{\varepsilon}_{ij.} - \overline{\varepsilon}_{i..} - \overline{\varepsilon}_{.j.} + \overline{\varepsilon}_{...})^2$$
$$= \sum_{i,j}\gamma_{ij}^2 + \sum_{i,j}(\overline{\varepsilon}_{ij.} - \overline{\varepsilon}_{i..} - \overline{\varepsilon}_{.j.} + \overline{\varepsilon}_{...})^2 + 2\sum_{i,j}\gamma_{ij}(\overline{\varepsilon}_{ij.} - \overline{\varepsilon}_{i..} - \overline{\varepsilon}_{.j.} + \overline{\varepsilon}_{...})$$

$$E\left[\frac{1}{n}SSAB\right] = \sum_{i,j}\gamma_{ij}^2 + \sum_{i,j}E(\overline{\varepsilon}_{ij.} - \overline{\varepsilon}_{i..} - \overline{\varepsilon}_{.j.} + \overline{\varepsilon}_{...})^2$$

$$
\begin{aligned}
E(\overline{\varepsilon}_{ij.} - \overline{\varepsilon}_{i..} - \overline{\varepsilon}_{.j.} + \overline{\varepsilon}_{...})^2 &= \mathrm{Var}(\overline{\varepsilon}_{ij.} - \overline{\varepsilon}_{i..} - \overline{\varepsilon}_{.j.} + \overline{\varepsilon}_{...}) \\
&= \mathrm{Var}(\overline{\varepsilon}_{ij.}) + \mathrm{Var}(\overline{\varepsilon}_{i..}) + \mathrm{Var}(\overline{\varepsilon}_{.j.}) + \mathrm{Var}(\overline{\varepsilon}_{...}) \\
&\quad - 2\mathrm{Cov}(\overline{\varepsilon}_{ij.}, \overline{\varepsilon}_{i..}) - 2\mathrm{Cov}(\overline{\varepsilon}_{ij.}, \overline{\varepsilon}_{.j.}) + 2\mathrm{Cov}(\overline{\varepsilon}_{ij.}, \overline{\varepsilon}_{...}) \\
&\quad + 2\mathrm{Cov}(\overline{\varepsilon}_{i..}, \overline{\varepsilon}_{.j.}) - 2\mathrm{Cov}(\overline{\varepsilon}_{i..}, \overline{\varepsilon}_{...}) - 2\mathrm{Cov}(\overline{\varepsilon}_{.j.}, \overline{\varepsilon}_{...})
\end{aligned}
$$

From the previous results, we know the formula for $\mathrm{Var}(\overline{\varepsilon}_{i..})$, $\mathrm{Var}(\overline{\varepsilon}_{.j.})$, $\mathrm{Var}(\overline{\varepsilon}_{...})$, $\mathrm{Cov}(\overline{\varepsilon}_{i..}$, $\overline{\varepsilon}_{...})$ and $\mathrm{Cov}(\overline{\varepsilon}_{.j.}, \overline{\varepsilon}_{...})$. Here we find the others

$$
\begin{aligned}
\mathrm{Var}(\overline{\varepsilon}_{ij.}) &= \mathrm{Var}\left(\sum_{k=1}^{n} \varepsilon_{ijk}/n\right) \\
&= \frac{1}{n^2}\left[\sum_{k} \mathrm{Var}(\varepsilon_{ijk}) + \sum_{k\neq t} \mathrm{Cov}(\varepsilon_{ijk}, \varepsilon_{ijt})\right] \\
&= \frac{1}{n}\sigma^2
\end{aligned}
$$

$$
\begin{aligned}
\mathrm{Cov}(\overline{\varepsilon}_{ij.}, \overline{\varepsilon}_{i..}) &= \mathrm{Cov}\left(\overline{\varepsilon}_{ij.}, \sum_{t=1}^{b} \overline{\varepsilon}_{it.}/b\right) \\
&= \frac{1}{b}\mathrm{Cov}(\overline{\varepsilon}_{ij.}, \overline{\varepsilon}_{ij.}) + \frac{1}{b}\sum_{t\neq j} \mathrm{Cov}(\overline{\varepsilon}_{ij.}, \overline{\varepsilon}_{it.}) \\
&= \frac{1}{b}\mathrm{Var}(\overline{\varepsilon}_{ij.}) + \frac{1}{b}\sum_{t\neq j} \mathrm{Cov}\left[\sum_{k_1}^{n}(\varepsilon_{ijk_1}/n), \sum_{k_2}^{n}(\varepsilon_{itk_2}/n)\right] \\
&= \frac{1}{bn}\sigma^2 + \frac{1}{bn^2}\sum_{t\neq j}\sum_{k} \mathrm{Cov}(\varepsilon_{ijk}, \varepsilon_{itk}) \\
&= \frac{1}{bn}\sigma^2 + \frac{1}{bn^2}(b-1)n\mathrm{Cov}(\varepsilon_{ijk}, \varepsilon_{itk}) \\
&= \frac{1}{bn}[\sigma^2 + (b-1)\sigma_{i12}]
\end{aligned}
$$

$$\begin{aligned}\mathrm{Cov}(\overline{\varepsilon}_{ij.},\overline{\varepsilon}_{.j.}) &= \mathrm{Cov}\left(\overline{\varepsilon}_{ij.},\sum_{t=1}^{a}\overline{\varepsilon}_{tj.}/a\right)\\ &= \frac{1}{a}\mathrm{Cov}(\overline{\varepsilon}_{ij.},\overline{\varepsilon}_{ij.}) + \frac{1}{a}\mathrm{Cov}\left(\overline{\varepsilon}_{ij.},\sum_{t\neq i}\overline{\varepsilon}_{tj.}\right)\\ &= \frac{1}{an}\sigma^2\end{aligned}$$

$$\begin{aligned}\mathrm{Cov}(\overline{\varepsilon}_{ij.},\overline{\varepsilon}_{...}) &= \mathrm{Cov}\left(\overline{\varepsilon}_{ij.},\frac{1}{a}\sum_{t=1}^{a}\overline{\varepsilon}_{t..}\right)\\ &= \frac{1}{a}\mathrm{Cov}(\overline{\varepsilon}_{ij.},\overline{\varepsilon}_{i..}) + \frac{1}{a}\mathrm{Cov}(\overline{\varepsilon}_{ij.},\sum_{t\neq i}\overline{\varepsilon}_{t..})\\ &= \frac{1}{abn}\left[\sigma^2 + (b-1)\sigma_{i12}\right]\end{aligned}$$

$$\begin{aligned}\mathrm{Cov}(\overline{\varepsilon}_{i..},\overline{\varepsilon}_{.j.}) &= \mathrm{Cov}\left(\overline{\varepsilon}_{i..},\frac{1}{a}\sum_{t=1}^{a}\overline{\varepsilon}_{tj.}\right)\\ &= \frac{1}{a}\mathrm{Cov}(\overline{\varepsilon}_{i..},\overline{\varepsilon}_{ij.}) + \frac{1}{a}\mathrm{Cov}\left(\overline{\varepsilon}_{i..},\sum_{t\neq i}\overline{\varepsilon}_{tj.}\right)\\ &= \frac{1}{abn}\left[\sigma^2 + (b-1)\sigma_{i12}\right]\end{aligned}$$

$$\begin{aligned}& E(\overline{\varepsilon}_{ij.} - \overline{\varepsilon}_{i..} - \overline{\varepsilon}_{.j.} + \overline{\varepsilon}_{...})^2\\ = \quad & \frac{(b-1)}{abn}\left[(a-1)\sigma^2 - (a-2)\sigma_{i12} - \frac{1}{a}\sum_{i}\sigma_{i12}\right]\end{aligned}$$

$$\sum_{i=1}^{a}\sum_{j=1}^{b} E(\overline{\varepsilon}_{ij.} - \overline{\varepsilon}_{i..} - \overline{\varepsilon}_{.j.} + \overline{\varepsilon}_{...})^2$$
$$= \frac{(a-1)(b-1)}{n}\left[\sigma^2 - \frac{1}{a}\sum_{i}\sigma_{i12}\right]$$

Therefore,

$$E\left[SSAB\right] = n\sum_{i,j}\gamma_{ij}^2 + (a-1)(b-1)\left[\sigma^2 - \frac{1}{a}\sum_{i}\sigma_{i12}\right]$$

and

$$E\left[MSAB\right] = \frac{n}{(a-1)(b-1)}\sum_{i,j}\gamma_{ij}^2 + \sigma^2 - \frac{1}{a}\sum_{i}\sigma_{i12}$$

Finally, we derive the formular for $E\left[MSE\right]$

$$E\left[SSE\right] = \sum_{i}\sum_{j}\sum_{k} E(\varepsilon_{ijk} - \overline{\varepsilon}_{ij.})^2$$

$$E(\varepsilon_{ijk} - \overline{\varepsilon}_{ij.})^2 = \mathrm{Var}(\varepsilon_{ijk}) + \mathrm{Var}(\overline{\varepsilon}_{ij.}) - 2\mathrm{Cov}(\varepsilon_{ijk}, \overline{\varepsilon}_{ij.})$$

$$\begin{aligned}\mathrm{Cov}(\varepsilon_{ijk}, \overline{\varepsilon}_{ij.}) &= \mathrm{Cov}\left[\varepsilon_{ijk}, \sum_{t=1}^{n}\varepsilon_{ijt}/n\right)\\ &= \frac{1}{n}\mathrm{Var}(\varepsilon_{ijk})\\ &= \frac{1}{n}\sigma^2\end{aligned}$$

$$E(\varepsilon_{ijk} - \overline{\varepsilon}_{ij.})^2 = \sigma^2 + \frac{1}{n}\sigma^2 - \frac{2}{n}\sigma^2$$
$$= \frac{n-1}{n}\sigma^2$$

Therefore,

$$E\left[SSE\right] = abn\frac{n-1}{n}\sigma^2$$

and

$$E\left[MSE\right] = \sigma^2$$

Appendix B

Derive the Covariance of Substitution Rates

Recalling Taylor's expansion:

$$f(x) \approx f(x_0) + f'(x_0)(x - x_0).$$

If $f(x) = \ln(x)$ then

$$\ln(x) \approx \ln(x_0) + \frac{1}{x}(x - x_0).$$

Therefore, replacing x with $\hat{\mu_1}$ and x_0 with μ_1

$$\ln(\hat{\mu}_1) \approx \ln(\mu_1) + \frac{1}{\mu_1}(\hat{\mu}_1 - \mu_1).$$

Simiarly,

$$\ln(\hat{\mu}_2) \approx \ln(\mu_2) + \frac{1}{\mu_2}(\hat{\mu}_2 - \mu_2).$$

Therefore,

$$\mathrm{Cov}(\ln \hat{\mu}_1, \ln \hat{\mu}_2) \approx \frac{1}{\mu_1 \mu_2}\mathrm{Cov}(\hat{\mu}_1, \hat{\mu}_2).$$

List of References

[1] Adachi, J. and M. Hasegawa. 1995. Tempo and mode of synomymous substitution s in mitochondrial DNA of primates. *Mol. Biol. Evol.* 13:200-208.

[2] Agresti, A. 1990. *Categorical data analysis.* Wiley, New York.

[3] Arnason, U., and A. Gullberg. 1993. Comparison between the complete mtDNA sequences of the blue and the fin whale, two species that can hybridize in nature. *J. Mol. Evol.* 37(4):312-322.

[4] Arnason, U., A. Gullberg, and B. Widegren. 1991. The complete nucleotide sequence of the mitochondrial DNA of the fin whale, Balaenoptera physalus. *J. Mol. Evol.* 33(6):556-568.

[5] Arnason, U., X. Xu, and A. Gullberg. 1996. Comparison between the complete mitochondrial DNA sequences of Homo and the common chimpanzee based on nonchimeric sequences. *J. Mol. Evol.* 42(2):145-152.

[6] Beard, C. B., D. M. Hamm., and F. H. Collins. 1993. The mitochondrial genome of the mosquito Anopheles gambiae: DNA sequence, genome organization, and comparisons with mitochondrial sequences of other insects. *Insect Mol. Biol.* 2(2):103-124.

[7] Billington, N. 2002. Mitochondrial DNA. Chapter 4 in *Population Genetics: Principles and Practices for Fisheries Scientists*(ed. E. Hallerman) American Fisheries Society.

[8] Bohr, V. A., and R. M. Anson. 1999. Mitochondrial DNA repair pathways. *J Bioenerg biomembr* 31:391-398.

[9] Brown, W. M., M. George, and A. C. Wilson. 1979. Rapid evolution of animal mitochondrial DNA. *Proc. Natl. Acad. Sci. USA* 76:1967-1971.

[10] Bulmer, M. 1989. Estimating the variability of substitution rates. *Genetics* 123:615-619.

[11] Bulmer, M. 1991. Use of the method of generalized least squares in reconstruction phylogenies from sequence data. *Mol. Biol. Evol.* 8(6):868-883.

[12] Bulmer, M., K. H. Wolfe, and P. M. Sharp. 1991. Synonymous nucleotide substitution rates in mammalian genes: implications for the molecular clock and the relationship of mammalian orders. *Proc. Natl. Acad. USA* 89:7844-7848.

[13] Cann, R. L., M. Stoneking, and A. Wilson. 1987. Mitochondrial DNA and human evolution. *Nature* 325:31-36.

[14] Casella, G., and R. L. Berger. 1990. Statistical Inference. Duxbury Press, Belmont, California.

[15] Chalmers, R. M. 2002. Mitochondrial myopathy, parkinsonism and multiple mtDNA deletions in a Sephardic Jewish family. *Neurology.* 58(4):670

[16] Chang, Y-S., F-L. Huang, and T-B. Lo. 1994. The complete nucleotide sequence and gene organization of carp (Cyprinus carpio) mitochondrial genome. *J. Mol. Evol.* 38(2):138-155.

[17] Dayoff, M. O., R. M. Schwartz, and B. C. Orcutt. 1978. A model of evolutionary change in proteins. pp 345-352 in *Atlas of Protein Sequence and Structure,* Vol. 5, Suppl. 3, edited by M. O. Dayhoff. National Biomedical Research Fooundation, Washington, D. C.

[18] Diggle, P. J., K. Liang, and S. L. Zeger. 1996. *Analysis of Longitudinal Data.* Oxford University Press Inc., New York.

[19] Efron, B., and R. J. Tibshirani. 1993. *An Introduction to the Bootstrap.* Chapman & Hall, New York.

[20] Eyre-walker, A., and B. S. Gaut. 1997. Correlated substitution rates among plant genomes. *Mol. Biol. Evol.* 14:455-460.

[21] Felsenstein, J. 1981. Evolutionary trees from DNA sequences: a maximum likelihood approach. *Mol. Biol. Evol.* 17:368-376.

[22] Fos, M., M .A. Dominguez, A. Latorre, and A. Moya. 1990. Mitochondrial DNA evolution in experimental populations of Drosophila subobscura. *Proceedings of the National Academy of Sciences* 87:4198-4201.

[23] Gaut, B. S., L. G. Clark, J. F. Wendel, and S. V. Muse. 1997. Comparisons of the molecular evolutionary process at *rbcL* and *ndhF* in the grass family (Poaceae). *Mol. Biol. Evol.* 14:769-777.

[24] Gaut, B. S., and M. T. Clegg. 1993. Nucleotide polymorphism in the ADH1 locus of pearl-millent *Peenisetum-Glaucum*) (*Aoaceae*) *Genetics* 135(4):1091-1097.

[25] Gaut, B. S., S. V. Muse, W. D. Clark, and M. T. Clegg. 1992. Relative rates of nucleotide substitution at the *rbcL* locus of Monocotyledonous plants. *J. Mol. Evol.* 35:292-303.

[26] Gaut, B. S., S. V. Muse, and M. T. Clegg. 1993. Relative rates of nucleotide substitution in the chloroplast genome. *Mol. Phylo. Evol.* 2:89-96.

[27] Green, D. R., and J. C. Reed. 1998. Mitochondria and apoptosis. *Science* 281: 1309-1312.

[28] Hasegawa, M., H. Kishino, and N. Saitou. 1985. Dating of the human-ape splitting by a molecular clock of mitochondrial DNA. *J. Mol. Evol.* 22:160-174.

[29] Hasegawa M., Kishino H., and Saitou N. 1991. On the maximum likelihood method in molecular phylogenetics. (Letter to the Editor). *J. of Mol. Evol.* 32:443-445.

[30] Hilton, H., A. Eyre-Walker, S. V. Muse, and B. S. Gaut. 1996. A survey of variation in animal mitochondria nucleotide substitution rates. unpublised paper.

[31] Horai, S., Y. Satta, K. Hayasaka, R. Kondo., T. Inoue, T. Ishida, S. Hayashi, and N. Takahata. 1992. Man's place in Hominoidea revealed by mitochondrial DNA genealogy. J. Mol. Evol. 35(1):32-43.

[32] Huelsenbeck, J. P. 1997. Phylogeny estimation and hypothesis testing using maximum likelihood. *Annu. Rev. Ecol. Syst.* 28:437-466.

[33] Huelsenbeck,J. P., B. Larget, and D. Swofford. 2000. A compound Poisson process for relaxing the molecular clock. *Genetics* 154:1879-1892.

[34] Jukes, T. H., and C. R. Cantor. 1969. Evolution of protein molecules. pp 21-23. In: H. M. Munro(ed) *Mammalian protein metabolism.* Academic Press, New York.

[35] Kavanagh, M. 1984. *A complet guid to monkeys, apes and other primates.* Viking Press, New York.

[36] Kendall M, and A. Stuart. 1979. *The Advanced Theory of Statistics.* Charles Griffin, London.

[37] Kimura, M. 1968. Genetics variability maintained in a finite population due to mutational production of neutral and early neutral isoalleles. *Genet. Res.* 11:247-269.

[38] Kimura, M. 1980. A simple method for estimating evolutionary rates of base substitutions through comparative studies of nucleotide sequences. *J. Mol. Evol.* 16:111-120.

[39] Kimura, M., and T. Ohta. 1972. Estimating the variability of substitution rates. *Genetics* 123:615-619.

[40] Kimura, M., and T. Ohta. 1974. On some principles governing molecular evolution. *Proc. Acad. Sci. USA* 71:2848-2852.

[41] King, J. L., and T. H. Juke. 1969. Non-Darwinian Evolution: random fixation of selectively neutral mutations. *Science* 164:788-798.

[42] Kishino, H., J. L. Thorne, and W. J. Bruno. 2001. Performance of a divergence time estimation method under a probabilistic model of rate evolution. *Mol. Bio. Evol.* 18(3): 352-361.

[43] Kocher, T. D., and A. C. Wilson. 1991. Sequence evolution of mitochondrial DNA in human and chimpanzees:control region and a protein-coding region. pp 391-413 in S.Osawa and T. Honjo, eds. *Evolution of life: fossils, molecules and culture.* Springer, Tokyo.

[44] Kumar, R., L. MarechalDrouard, K. Akama, and I. Small. 1996. Striking differences in mitochondrial tRNA import between different plant species. *Molecular & General Genetics.* 252(4):404-411.

[45] Ladoukakis, E. D., and E. Zouros. 2001. Direct evidence for homologous recombination in mussel (Mytilus galloprovincialis) mitochondrial DNA. *Mol. Biol. Evol.* 18(7):1168-1175.

[46] Ladoukakis, E. D., and E. Zouros. 2001. Recombination in animal mitochondrial DNA: evidence from published sequences. *Mol. Bio. Evol.* 18(11):2127-2131.

[47] Lewis, D. L., C. L. Farr, and L. S. Kaguni. 1995. Drosophila melanogaster mitochondrial DNA: completion of the nucleotide sequence and evolutionary comparisons. *Insect Mol. Biol.* 4(4):263-278.

[48] Li, P., J. Bousquest. 1992. Relative-rate test for nucleotide substitutions between two lineages. *Mol. Biol. Evol.* 9:1185-1189.

[49] Li, W-H. and C-I. Wu. 1985. Evidence for higher rates of nucleotide substitution in rodents than in man. *Proc. Natl. Acad. Sci.* USA. 82:1741-1745.

[50] Li, W-H., C-I. Wu, and C-C. Luo. 1985. A new method for estimating synonymous and nonsynonymous rates of nucleotide substitution considering the relative likelihood of nucleotide and codon changes. *Mol. Biol. Evol.* 2:150-174.

[51] Li, W-H., D. L. Ellsworth, J. Krushkal, B. H-J. Chang, and D. Hewett-emmet. 1996. Rates of nucleotide substitution in primates and rodents and the generation-time effect hypothesis. *Mol. Phylo. Evol.* 5:182-187.

[52] Liang, K. Y., and S. L. Zeger. 1986. Longitudinal data analysis using generalized linear models. *Biometrika* 73:13-22.

[53] Lopez, J. V., S. Cevario, and S. J. O'Brien. 1996. Complete nucleotide sequences of the domestic cat (Felis catus) mitochondrial genome and a transposed mtDNA tandem repeat (Numt) in the nuclear genome. *Genomics* 33(2):229-246.

[54] Jin, L., and M. Nei. 1990. Limitations of the evolutionary parsimony method of phylogenetic analysis. *Mol. Biol. Evol.* 7(1):82-102.

[55] MacRae, A. F., and W. W. Anderson. 1988. Evidence for non-neutrality of mitochondrial DNA haplotypes in Drosophila pseudoobscura. *Genetics* 120:485-494.

[56] Malhotra, A., and R. S. Thorpe. 1994. Parallels between island lizards suggests selection on mitochondrial DNA and morphology. *Proceedings of the Royal Society of London (Series B)* 257:37-42.

[57] Martin, A. P., Naylor, G. J. P., and S. R. Palumbi. 1992. Rates of mitochondrial DNA evolution in sharks are slow compared with mammals. *Nature* 359:153-155.

[58] Martin, A. P., and S. R. Palumbi. 1993. Body size, metabolic rate, generation time, and the molecular clock. *Proc. Natl. Acad. Sci. USA* 90(9):4087-91.

[59] Menzies, F. M., P. G. Ince, P. J. Shaw. 2002. Mitochondrial involvement in amyotrophic lateral sclerosis. *Neurochem Int.* 40(6):543-551.

[60] Meunier, J., and A. Eyre-Walker. 2001. The correlation between linkage disequilibrium and distance: implications for recombination in hominid mitochondria. *Mol. Biol. Evol.* 18(11):2132-2135.

[61] Moran, N. A. 1996. Accelerated evolution and Muller's rachet in endosymbiotic bacteria. *Proc. Natl. Acad. Sci. USA* 93:2873-2878.

[62] Murphy, A. N., G. Fiskum, and M. F. Beal. 1999 Mitochondria in neurodegeneration: bioenergetic function in cell life and death. *J Cereb Blood Flow Metab* 19:231-245.

[63] Muse, S. V., and B. S. Gaut. 1994. A likelihood approach for comparing synonymous and nonsynonymous nucleotide substitution rates, with application to the chloroplast genome. *Mol. Bio. Evol.* 11(5):715-724.

[64] Muse, S. V., and B. S. Gaut. 1997. Comparing patterns of nucleotide substitution rates among chloroplast loci using the relative ratio test. *Genetics* 146:393-399.

[65] Muse, S. V., and B. S. Weir. 1992. Testing for equality of evolution rates. *Genetics* 132:269-276.

[66] Nei, M., and T. Gojobori. 1986. Simple methods for estimating the numbers of synonymous and nonsynonymous nucleotide substitutions. *Mol. Biol. Evol.* 3(5):418-426.

[67] Nei, M., J. C. Stephens, and N. Saitou. 1985. Methods for computing the standard errors of branceing points in an evolutionary tree and their application to molecular data from humans and apes. *Mol. Bio. Evol.* 2:66-85.

[68] Nelder, J. A. and R. W. M. Wedderburn. 1972. Generalized linear models. *J. Roy.Stat. Soci.* A 135:370-384.

[69] Neter, J., W. Wasserman, and M. Kutner. 1990. *Applied Linear Models: Regression, Analysis of Variance and Experimental Design.* Richard D. Irwin, Inc.

[70] arsons, T. J., D. S. Muniec, K. Sullivan, N. Woodyatt, R. Alliston-Greiner, M. R. Wilson, D. L. Berry, K. A. Holland, V. W. Weedn, P. Gill, and M. M. Holland. 1997. A high observed substitution rate in the human mitochondrial DNA control region. *Nature Genetics* 15(4):363-367.

[71] Press, W. H., S. A. Teukolsky, W. T. Vetterling, and B. P. Flannery. 1992. *Numerical Recipes in C.* Cambridge University Press.

[72] Rustin, R., J. C. von Kleist-Retzow, Z. Vajo, A. Rtig, and A. Munnich. 2000. For debate: defective mitochondria, free radicals, cell death, aging-reality or myth-ochondria? *Mechan. Age Dev.* 114: 201-206.

[73] Saitou, N., and M. Nei. 1987. The neighbor-joining method: a new method for reconstruction phylogenetic trees. *Mol. Biol. Evol.* 4:406-425.

[74] Sanderson, M. J. 1997. A nonparametric approach to estimating divergence times in the absence of rate constancy. *Mol. Biol. Evol.* 14:1218-1232.

[75] Sanderson, M. 2002. Estimating absolute rates of molecular evolution and divergence times: a penalized likelihood approach. *Mol. Bio. Evol.* 19(1):101-109.

[76] Sarich, V. M., and A. C. Wilson. 1973. Generation time and genomic evolution in primates. *Science, USA* 58:142-148.

[77] Sokal, R. R., and F. J. Rohlf. 1995. *Biometry.* W. H. Freeman, New York.

[78] Stains, H. J. 1984. Carnivores. pp 491-521 in Anderson, S. and J. K. Jones, Jr. (eds). Orders and Families of Recent Mammals of the World. John Wiley and Sons, New York.

[79] Tamura, K. 1992. Estimation of the number of nucleotide substitutions when there are strong transition$-$transversion and $G + C-$content biases. *Mol. Bio. Evol.* 9:676-687.

[80] Tajima, F., and M. Nei. 1984. On the stochastic model for estimation of mutational distance between homologous proteins. *J. Mol. Evol.* 2:87-90.

[81] Tamura, K., and M. Nei. 1993. Estimation of the number of nucleotide substitutions in the control region of mitochondrial DNA in human and chimpanzees. *Mol. Biol. Evol.* 19:512-526.

[82] Tavaré, S. 1986. Some probabilistic and statistical problems in the analysis of DNA sequences. In: Miura RM (ed.). Lecture on mathematics in the Life Sciences, pp57-86. *American Mathematic Socitety,* Providence, RI.

[83] Thorne J. L., H. Kishino, and J. Felsenstein. 1992. Inching toward reality: an improved likelihood model of sequence evolution. *J. Mol. Evol.* 34:316.

[84] Thorne, J. L., H. Kishino, I. S. Painter. 1998. Estimating the rate of evolution of the rate of molecular evolution. *Mol. Biol. Evol.* 15(12):1647-1657.

[85] Tzeng C-S., C-F. Hui, and S-C. Shen, and P-C. Huang. 1992. The complete nucleotide sequence of the Crossostoma lacustre mitochondrial genome: conservation and variations among vertebrates. *Nucleic Acids. Res.* 20(18):4853-4858.

[86] Uzzell, T., and K. W. Corbin. 1971. Fitting discrete probability distributions to evolutionary events. *Science* 172:1089-1096.

[87] Weber, M. and L. F. DeBeaufort. 1916. *The fishes of the Indo-Australian Archipelago.* Leiden, The Netherlands.

[88] Weir B. S. 1990. Genetic Data Analysis. *Sinauer* Sunderland, MA.

[89] Wiedemann, F. R., G. Manfredi, C. Mawrin, M. F. Beal, E. A. Schon. 2002. Mitochondrial DNA and respiratory chain function in spinal cords of ALS patients. *J. Neurochem.* 80(4):616-625.

[90] Wolfe, K. H., W-H. Li, and P. M. Sharp. 1985. Rates of nucleotide substitution vary greatly among plant mitochondrial,chloroplast, and nuclear DNAs. *Proc. Natl. Acad. Sci., USA.* 84:9054-9058.

[91] Wu, C-I., and W-H. Li. 1985. Evidence for higher rates of nucleotide substitution in rodents than in man. *Proc. Natl. Acad. Sci., USA.* 82:1741-1745.

[92] Xu, X., A. Janke, and U. Arnason. 1996. The complete mitochondrial DNA sequence of the greater Indian rhinoceros, Rhinoceros unicornis, and the Phylogenetic relationship among Carnivora, Perissodactyla, and Artiodactyla (+ Cetacea). *Mol. Biol. Evol.* 13(9):1167-1173.

[93] Yang, Z., 1994. Estimating the pattern of nucleotide substitution. *J. Mol. Evol.* 39:105-111.

[94] Yang, Z. 1994. Maximum likelihood phylogenetic estimation from DNA sequences with variable rates over sites: approximate methods. *J. Mol. Evol.* 39:306-314.

[95] Zardoya R., A. Garrido-Pertierra, and J. M. Bautista. 1995. The complete nucleotide sequence of the mitochondrial DNA genome of the rainbow trout, Oncorhynchus mykiss. *J. Mol. Evol.* 41(6):942-951.

[96] Zurawski, G., and M. T. Cleff. 1987. Evolution of high-plant chloroplast encoded genes: implications for structure-function and phylogenetic studies. *Annu. Rev. Plant. Physiol.* 38:391-418.