CONTEXT DEPENDENT DNA SUBSTITUTION MODELS

ZHANG RONGLI

(Master of Science, National University of Singapore)

(Bachelor of Mathematics, Beijing Jiaotong University)

A THESIS SUBMITTED

FOR THE DEGREE OF DOCTOR OF PHILOSOPHY

DEPARTMENT OF STATISTICS AND APPLIED PROBABILITY

NATIONAL UNIVERSITY OF SINGAPORE

2009

Acknowledgements

This thesis would not have been possible without the support and help of many people. It is pleasant that I have now the opportunity to express my gratitude for all of them.

First of all, I would like to express my deep and sincere gratitude to my supervisor, Assistant Professor Yap Von Bing, whose continuous support and encouragement have been crucial to the completion of this thesis. He gives me a lot invaluable advice and guidance during my PhD study period. I truly appreciate all the time and effort he has spent in helping me to solve the problems I encountered. His patience and encouragement help me to overcome a lot of difficulties.

I am greatly indebted to teachers who have inspired and helped me to enter the field of statistics. Professor Bai Zhidong, Professor Chen Zehua are to be mentioned particularly. I also thank Professor Chua Ting Chiu , Professor Kuk Anthony and Professor Choi Kwok Pui for their kind support. I express my appreciation other members and staff of the Statistics department for their help in various ways and providing such a pleasant research environment.

It is a great pleasure to record my thanks to my dear friends, Ms Li Yue, Ms Zhao Jingyuan, Ms Hao Ying, Ms Wang Xiaoying and Ms Zhao Wanting, who have given me much help in my study and life. I also wish to express my gratitude to my friend Khang Tsung Fei for his kind support. I Sincerely thanks all my friends who helped me in one way or another and for taking caring of me and encouraging me.

I feel a deep sense of gratitude for my husband for his love, encouragement, support and understanding during the PhD period. I also thank my son for giving me love and happiness.

Finally, I would like to give my special thanks to my parents for their support and encouragement.

Contents

Summary

Independent substitution model study is a classical topic in molecular evolution. However, empirical evidence suggests that the context dependent model is a more accurate description of the DNA evolution process. Thus, there is a great demand for statistical approaches for context dependent substitution models, which can help better understand the evolution relationship of species.

In this thesis, we propose a general context dependent framework. Based on the framework, we investigate two-flanking sites context dependent model and derive two sub-models by clustering the substitution matrices. Moreover, we develop a modified parsimony method and maximum pseudo-likelihood method to estimate the parameters in our models. We conduct experiment on the simulation data for our proposed models and methods. The methods were also applied to the real data.

Our work is different from previous work in the following aspects:

(1)The problem: Previous works on context dependent models investigated the estimation of substitution rates from two known descendent sequences that evolved from the same unknown ancestor sequence. Little research was done to estimate context dependent substitution rates from a given ancestor sequence and its descendent sequence. In our work, the rate estimation was based on the evolution from a known ancestor to a known descendent. We made use of the phylogenetic tree of the species to first estimate the the ancestor.

(2)Model definition: We propose a general context dependent model framework, which used a mathematical of representation to describe the general cases of context dependent and independent models. Based on the general model, different context dependent models can be derived as the special cases of the general model.

(3)Model simplification: In context dependent substitution models, to describe the substitution process, substitution matrices are defined for different context. This inevitably introduces many parameters. The usual approach for reducing the number of parameters is to reduce the number of independent parameters in each substitution matrix. We have proposed to reduce the number of matrices based on the knowledge of DNA evolution. Simulation showed that our models work well. To reduce the number of matrices, the contexts need to be grouped together. In the thesis, we propose to use statistical method to cluster the context cases. This not only confirms our grouping methods but also provides a general way of handling this problem.

(4) Estimation methods: Parsimony approach is normally used in the estimation of independent substitution models. We have proposed an improved parsimony method and applied it to context dependent models. It overcomes the inaccuracy of usual meth-

ods in dealing with adjacent changes in DNA evolution. Experiment shows an improvement over the usual approach. We have proposed to use direct optimization of the pseudo-likelihood approach. However, optimization starting from a set of fixed initial values takes too long to converge. By providing the rates estimated from Parsimony method as the initial optimization values, the optimization process converges quickly.

(5) Simulation process and evaluation methods: Previous research normally worked with limited real data. In our work, we have developed a process to simulate context dependent DNA sequence evolutions. This provides us a flexibility of doing various experiment on simulated data. In the evaluation of different models, we have used the adjusted pseudo-likelihood ratio test.

List of Tables

List of Figures

Chapter 1

Introduction

The molecular evolution process is normally studied by looking at nucleotide substitutions in DNA sequence. Substitution is a process whereby a nucleotide changes from one state to another in a collection of populations. It is the result of mutation, selection and fixation (p53, Graur and Li. 2000). Substitution models are used to describe the process of nucleotide changes. Methods with different assumptions have been proposed to model the substitution process.

Most of the existing models for nucleotide substitution process assume that neighboring sites evolve independently. The independent assumption is just an approximation of the actual evolution process because it has been observed that neighboring nucleotides do have an effect on the substitution of nucleotides (Krawczak et al. 1998). Therefore, when dealing with substitution rates, we need to consider context dependent substitution models, which allow the substitution of nucleotides to depend on their neighboring nucleotides. In the following sections, the background knowledge about DNA evolution will be introduced and the literature related to context dependent models will be reviewed in detail.

1.1 DNA sequence

The hereditary information in an organism is carried by DNA (deoxyribonucleic acid) molecules. DNA usually consists of two complementary strands twisted around each other to form a double helix. Each strand is a linear polynucleotide consisting of four kinds of nucleotides: adenine(A), guanine(G), cytosine(C) and thymine(T). The four nucleotides are grouped into two purines (A and G) and two pyrimidines (C and T). The two complementary strands are joined through the pairing of complementary nucleotides. A always pairs with T, and G always pairs with C.

In molecular evolution, a sequence alignment is a way of arranging the sequences of DNA, RNA, or protein to identify regions of similarity that may be a consequence of functional, structural, or evolutionary relationships between the sequences. Aligned sequences of nucleotide or amino acid residues are typically represented as rows within a matrix. Gaps can be inserted into compared sequences so that identical or similar characters are aligned in successive columns.

In the typical case, a DNA sequence is represented by a string of letters, e.g.

$$AAAGTCTGAC,$$

in which each of the letter represents a nucleotide. When a substitution happens, one of the letters will change to other types. Sometimes insertion or deletion of nucleotides may happen during evolution. When an evolved sequence is aligned with its original sequence, the alignment may be punctuated by gaps. In this thesis, we disregard the gaps and consider only the point mutations where a single nucleotide is replaced by another nucleotide. We also do not consider the simultaneous substitution of more than one nucleotide at one time. Whealan and Goldman (2004) described a model which allows doublet and triplet mutations.

The information carried by DNA is held in the sequence of pieces of DNA called genes. A gene is a sequence of DNA that contains genetic information and can influence the phenotype of an organism. Within a gene, the sequence of bases along a DNA strand is transcribed into a messenger RNA sequence, which is then translated into amino acid. The relationship between the nucleotide sequences of genes and the amino-acid sequences of proteins is determined by the rules of translation, known collectively as the genetic code. The genetic code consists of three-letter "words" called codons (e.g. ACT, CAG, TTT). It is a set of rules whereby information encoded in genetic material (DNA or RNA sequences) is mapped into amino acid by the cellular machinery.

Since a codon consists of three nucleotides and there are four different types of nucleotides, there are $4^3 = 64$ possible codons. In the genetic code, 61 of these codons code for specific amino acids and are called nonstop codons; while the remaining three are stop codons. The stop codons are for the standard genetic code UAG (in RNA) /

TAG (in DNA) , UAA/TAA , and UGA/TGA. Translation stops when a stop codon is encountered. There are only 20 amino acids, so some of the 61 nonstop codons encode the same amino acid. Codons that map into the same amino acid are synonymous; otherwise they are nonsynonymous.

When we look at the DNA sequences of two species (an ancestor species and a descendant species), normally the length of the two sequences are different due to insertion and deletion of nucleotides in evolution. To determine the extent of similarity between them, the DNA sequences from two species have to be aligned first. The alignment identifies conserved regions, and divergent regions so that the phylogeny between a group of species can be inferred.

1.2 Markov processes

In analysis of DNA sequences, much of the mathematics of the nucleotide substitution process relies on the assumption of a stationary homogeneous Markov process (Kelly 1979). Briefly, we describe what this process is about.

Let $X(t)$ be a stochastic process taking values in a finite state space S for $t \in [0, \infty)$. If $(X(t_1), X(t_2), \ldots, X(t_n))$ has the same distribution as $(X(t_1 + s), X(t_2 + s), \ldots, X(t_n + s))$ for all $t_1, t_2, \ldots, t_n, s \in [0, \infty)$, then the stochastic process $X(t)$ is stationary. The stochastic process $X(t)$ is a Markov process if for any $n \geq 1$, and $0 \leq t_1 \leq t_2 \leq \ldots \leq$

$t_n \leq t_{n+1}$,

$$Pr(X(t_{n+1}) = j_{n+1}|X(t_1) = j_1, X(t_2) = j_2, \ldots, X(t_n) = j_n) = Pr(X(t_{n+1}) = j_{n+1}|X(t_n) = j_n) \tag{1.1}$$

for any $j_1, \ldots, j_{n+1} \in S$. In simple words, equation (1.1) says that, given the present state, the future and past states are independent.

A Markov process is time homogeneous if $Pr(X(t+s) = j|X(t) = i)$ does not depend on t. For a time homogeneous continuous time Markov process, $P(t) = \{p_{ij}(t)\}$ and

$$p_{ij}(t) = Pr(X(s+t) = j|X(t) = i) \tag{1.2}$$

for any s and t.

A stationary distribution π is a vector whose entries sum to 1, and satisfies the equation

$$\pi = \pi P(t) \tag{1.3}$$

for any t.

Let $X(t)$ be a homogeneous continuous time Markov process with a finite state space of four nucleotides. A Markov process is usually specified by a rate matrix Q, whose elements represent instantaneous substitution rates among the four nucleotides. The rate matrix Q is defined as follows:

(1) The transition rate from state i to state j ($i \neq j$) is defined as

$$q_{ij} = \lim_{s \to 0} \frac{Pr(X(t+s) = j|X(t) = i)}{s}. \tag{1.4}$$

(2) The diagonal entry q_{ii} is defined as

$$q_{ii} = -\sum_{j \neq i} q_{ij} \tag{1.5}$$

Let $P(t) = \left\{p_{ij}(t)\right\}$ be the transition probability matrix, that is,

$$p_{ij}(t) = Pr(X(t) = j|X(0) = i),$$

then $P(t)$ is given by

$$P'(t) = P(t)Q. \tag{1.6}$$

That is

$$P(t) = \exp(tQ). \tag{1.7}$$

1.3 Independent substitution models

Statistical models that deal with DNA sequence evolution can be constructed from individual nucleotides or codons. A standard assumption is that nucleotides along the DNA sequence evolve independently of one another. For codon models, it is normally assumed that the nucleotides within a codon are context dependent; the codons, however, are assumed to evolve independently of one another.

1.3.1 Nucleotide substitution models

In homologous DNA sequences, nucleotide substitution is commonly assumed to follow a stationary homogeneous Markov process. The rate matrix Q has at most $4^2 - 4 = 12$

parameters. The most general form of substitution model is the unrestricted model, in which there are no no constraints between parameters, i.e. all 12 parameters are free parameters, as shown in Q Matrix 1. The "-" symbols along the main diagonals indicate elements to be defined as $q_{i,i} = -\sum_{j:j\neq i,} q_{i,j}$.

Q matrix 1: Unrestricted substitution rate matrix

$$\begin{array}{c|cccc}
 & \mathrm{A} & \mathrm{C} & \mathrm{G} & \mathrm{T} \\
\hline
\mathrm{A} & - & a & b & c \\
\mathrm{C} & d & - & e & f \\
\mathrm{G} & g & h & - & i \\
\mathrm{T} & j & k & l & -
\end{array}$$

Constraints can be imposed to reduce the number of free parameters while still retaining sufficient accuracy.

Tavare (1986) first proposed the reversible substitution models. A stationary Markov process $X(t)$ is reversible if and only if there exists a collection of positive numbers π_j summing to unity that satisfy the balanced equations

$$\pi_i q_{ij} = \pi_j q_{ji} \tag{1.8}$$

where $1 \leq i, j \leq 4$. If this condition holds, then π is the stationary distribution of the process, and the reversible model can be obtained. It reduces the number of free parameters to 9, as shown in Q matrix 2. They assumed that $q_{ij} = a_{ij}\pi_j$, then from the equation (1.8), we can obtain

$$\pi_i a_{ij} \pi_j = \pi_j a_{ji} \pi_i, \tag{1.9}$$

that is, $a_{ij} = a_{ji}$. For example, $a_{12} = a_{21} = a$, $a_{13} = a_{31} = b$ and so on.

Q matrix 2: The general reversible substitution rate matrix

$$\begin{array}{c|cccc} & \text{A} & \text{C} & \text{G} & \text{T} \\ \hline \text{A} & - & a\pi_C & b\pi_G & c\pi_T \\ \text{C} & a\pi_A & - & d\pi_G & e\pi_T \\ \text{G} & b\pi_A & d\pi_C & - & f\pi_T \\ \text{T} & c\pi_A & e\pi_C & f\pi_G & - \end{array}$$

One widely-used model is the HKY model (Hasegawa, Kishino and Yano 1985), as shown in Q matrix 3. The HKY model has a parameter κ for the ratio of the rates of transition (a change within A, G or within C, T) to transversion (a change from one of the groups A, G and C, T to the other), and allows for a general stationary distribution $\pi = (\pi_A, \pi_G, \pi_C, \pi_T)$ of the Markov process. There are altogether five parameters.

Q matrix 3: The HKY85 substitution rate matrix

$$\begin{array}{c|cccc} & \text{A} & \text{C} & \text{G} & \text{T} \\ \hline \text{A} & - & \pi_C & \kappa\pi_G & \pi_T \\ \text{C} & \pi_A & - & \pi_G & \kappa\pi_T \\ \text{G} & \kappa\pi_A & \pi_C & - & \pi_T \\ \text{T} & \pi_A & \kappa\pi_C & \pi_G & - \end{array}$$

If $\pi = (\frac{1}{4}, \frac{1}{4}, \frac{1}{4}, \frac{1}{4})$, the model reduces to the JC69 model (Jukes and Cantor 1969), which is the earliest and simplest model. In this model, all nucleotides undergo transitions at the same rate.

1.3.2 Codon substitution models

Substitution models for independent codon sequence are much more complicated. On one hand we should keep some of the modeling ideas from the nucleotide models; on the other hand, we need to take into consideration translation of a codon to its corresponding amino acid. Because of differences in their effects on the physiology of an organism, synonymous and nonsynonymous substitutions can have quite different dynamics. For example, synonymous substitutions usually occur at a much faster rate than nonsynonymous substitutions. Hence, in coding sequences it is often desirable to separate these two.

We assume that mutations occur at the three codon positions independently, and only single-nucleotide substitutions are permitted to occur instantaneously, as mutations involving more than one position will be ignored. The evolutionary processes in the codons are assumed to be independent identical Markov processes with rates described by a matrix with 61×61 entries.

DNA substitution mutations are of two types. Transitions are interchanges of $A \leftrightarrow G$, $C \leftrightarrow T$. Transversions are interchanges of $A \leftrightarrow T$, $G \leftrightarrow T$, $A \leftrightarrow C$ and $C \leftrightarrow G$.

Goldman and Yang (1994) proposed a complex model that incorporates a transition/transversion parameter and differentiates different nonsynonymous changes. They considered different synonymous (d_S) and nonsynonymous (d_N) substitution rates. Yang (1998) developed the codon-based likelihood models that allow for variable d_N/d_S ratios among lineages. Following the notation of Yang (1998), the rate matrix is termed Q, and individual entries in this matrix, termed q_{ij}, correspond to the relative rate of change from codon i to codon j. The $q_{ij}(i \neq j)$ are defined as

$$q_{ij} = \begin{cases} 0 & \text{more than one nucleotide difference} \\ \pi_j & \text{synonymous transversion} \\ \pi_j \kappa & \text{synonymous transition} \\ \pi_j \omega & \text{nonsynonymous transversion} \\ \pi_j \kappa \omega & \text{nonsynonymous transition.} \end{cases}$$

where κ is the transition/transversion rate ratio, ω is the nonsynonymous/synonymous rate ratio, and π_j is the equilibrium frequency of codon j, calculated from the nucleotide frequencies at the three codon positions. Under this model, $\omega = d_N/d_S$.

1.4 Context dependent substitution models

The independent model is a crude approximation in many cases because change of nucleotides is actually affected by its neighboring sites in real data, i.e. the CpG effect where an excess of $C \rightarrow T$ substitutions is observed at positions with a CpG dinucleotide (Gojobori et al. 1982). Ideally we have to consider the context of the sites in substitution model. Therefore, recently neighboring dependence has been considered in substitution models. Context dependent substitution models describe this kind of substitution process. Recently, a lot of mathematical and computational frameworks have been introduced to construct the context dependent substitution models. Arndt et al.(2003a) and Arndt and Hwa (2005) considered the case where the ancestral sequence is known. Lunter and Hein (2004) and Hwang and Green (2004) considered an unknown ancestral sequence. Christensen (2006) proposed the sequence distribution at the root for the unknown ancestor.

Various methods have been proposed for the special case of two sequences and a reversible substitution process that allows for general context-dependent substitution, with substitution rates for each base depending on the identity of flanking bases. These models reflect more accurately an assumed process of context-dependent substitution. With these models, the likelihood computation can no longer be expressed as a product over the sites of an alignment, and exact parameter estimation becomes intractable. Markov chain Monte Carlo (MCMC) (Lunter and Hein 2004; Hwang and Green 2004) and Expectation-Maximization (EM) (Christensen 2006) algorithms are needed for parameter estimation.

1.4.1 Context dependent model at the nucleotide level

1. Mixture model

Arndt et al.(2003a) considered a context dependent model at the nucleotide level suitable for the description of the noncoding parts of the genome. They derived an approximation to the stationary distribution as follows.

Let $\lambda(y_i|x_{i-1}, x_i, x_{i+1})$ be the rate for a change of x_i to y_i, when the two neighboring nucleotides are x_{i-1} and x_{i+1}. It is modeled linearly as:

$$\lambda(y_i|x_{i-1}, x_i, x_{i+1}) = \lambda_0(y_i|x_i) + \lambda_l(y_i|x_{i-1}, x_i) + \lambda_r(y_i|x_i, x_{i+1}) \tag{1.10}$$

where λ_0 is a rate not depending on the context, λ_l is a rate depending on the left neighbor, and λ_r is a rate depending on the right neighbor.

Arndt et al.(2003b) used the model from Arndt et al. (2003a) with four parameters in λ_0, one nonzero term in λ_l ($CG \to CA$), and one nonzero term in λ_r ($CG \to TG$). Assuming the ancestor is known, they used the pseudo-likelihood instead of calculating the true likelihood under the model. The likelihood is approximated by a product of marginal likelihood of the form $P(x_i(T)|x_{i-1}(0), x_i(0), x_{i+1}(0))$ for state T, where $T \in (A, C, G, T)$.

Arndt and Hwa (2005) defined a substitution model which included all neighbor-independent single nucleotide changed and additional neighbor-dependent processes. Based on this substitution model, they estimated the relative substitution frequencies and judged their importance in order to be included into the modeling. To estimate the substitution frequencies, the authors compared a pair of ancestral sequence $x = (x_1 x_2 \ldots x_n)$ and its daughter sequence $y = (y_1 y_2 \ldots y_n)$, where the daughter sequence represents the state of the ancestral sequence after the latter has undergone substitution processes for some time.

The log likelihood for sequence y evolving from ancestral sequence x under a given substitution model parameterized by the substitution frequencies $\{r\}$ is given by

$$\begin{aligned}\log L_{\{r\}} &= \log P_{\{r\}}(y|x) \\ &\approx \log \prod_{i=2}^{L-1} P_{\{r\}}(y_i|x_{i-1}, x_i, x_{i+1}) \\ &= \sum_{x_1x_2x_3} n(x_1x_2x_3 \to y_2) \log P_{\{r\}}(y_2|x_1x_2x_3)\end{aligned} \tag{1.11}$$

where $P_{\{r\}}(y|x)$ is the probability of the evolution of the sequence x into y. The numbers $n(x_1x_2x_3 \to y_2)$ denote the counts of observations of a base substitution from x_2 (flanked by x_1 to the left and x_3 to the right) to y_2.

2. Overlapping dinucleotide substitution model

Lunter and Hein (2004) introduced the over-lapping dinucleotide substitution model which allows only single nucleotide substitutions. They considered the neighbor pair sites together. Since there are four different types of nucleotides, there are $4^2 = 16$ possible pairs. Thus, the parameters of the model are given by a 16×16 rate matrix M. These rates apply to each of the $L - 1$ pairs of neighboring nucleotides in a sequence of length L simultaneously. The matrix R_k has dimension $4^L \times 4^L$, and corresponds to M acting on nucleotides k and $k + 1$ only, with no mutation process acting on any other nucleotides. The full model has rate matrix $R = \sum_{k=1}^{L-1}$, corresponding to the dinucleotide substitution process acting on all $L - 1$ di-nucleotides simultaneously.

For the substitution model, they used only a subset of the 240 free parameters in the matrix M. The symmetry of the substitution process under reverse-complement means that all mononucleotide substitutions can be described by the $4 \times 4 \times 3 = 48$ right-neighbour rates only. They used a single dinucleotide substitution rate with 49 parameter in all in their analysis.

They also derived an algorithm to calculate the likelihood of observing sequences evolving under this model. They used Bayesian MCMC sampling to infer the model parameters. In their approach, they used a recursive algorithm for approximation of likelihood function.

3. Two flanking nucleotides substitution model

Let $x = (x_1, x_2, \ldots, x_n)$ be a DNA sequence, where x_i is either a single nucleotide or a single codon, and let $x(t)$ be the process at time t. Here, we review papers where the rate of change of x_i depends on its two flanking neighbors: x_{i-1} and x_{i+1}. Such models are known as being context dependent on the two flanking nucleotides.

Hwang and Green (2004) described a context dependent model which allows the substitution rate at each site to depend on the two flanking nucleotides. For example, consider two sequence:

Seq1: $AACTAGTGA$

Seq2: $ACGAGCATA$

The two rates of $T \rightarrow A$ are the same in independent case. But in context dependent case, the two rates between the $T \rightarrow A$ are different. The 4th position $T \rightarrow A$ depend on CA, the 7th position $T \rightarrow A$ depend on GG.

In the independent case, we use one 4×4 substitution matrix to describe the substitution process. In the context dependent model, we use 16 of 4×4 substitution rate matrice to describe their evolution process.

Hwang and Green (2004) assumed that the model is nonstationary and they used a second order Markov chain model for the distribution of the common ancestor sequence of the observed sequences. Their context dependent model allows the substitution rate at each site to depend on the two flanking nucleotides. That is, each site is dependent on their left and right neighboring sites. A Bayesian MCMC approach was used to obtain samples from the posterior distribution of the parameters. The authors used a discrete time approximation of the substitution process for inference in their MCMC approach.

Christensen (2006) extended Christensen et al.'s (2005) work to the nonreversible and nonstationary nucleotide substitution models. The author also constructed a pseudo-likelihood method for inference in nonreversible nucleotide substitution models with neighbor dependent substitution rates. Maximization of the pseudo-likelihood was done using the EM algorithm.

Hobolth(2008) described statistical inference of neighbor-dependent models using a Markov chain Monte Carlo expectation maximization (MCMC-EM) algorithm.

4. Phylogenetic model

Siepel and Haussler (2004) introduced methods for incorporating context-dependent substitution into phylogenetic models. They considered N-tuples of nucleotides, where N is either 1, 2 or 3. There are three properties of their model. First, its characterization of context-dependent substitution within N-tuples of adjacent sites is explicit. Second, it is able to accommodate overlapping N-tuples. Third, the parameterization of the substitution process is rich.

For nonoverlapping N-tuples, the parameters were estimated using an EM algorithm, with a quasi-Newton algorithm for the maximization step. Overlapping N-tuples were efficiently handled by assuming Markov dependence of the observed bases at each site on those at the $N-1$ preceding sites, and the required conditional probabilities were computed using an extension of Felsenstein's algorithm (Felsenstein 1981b).

1.4.2 Codon context substitution models

If the rate of a change for a site depends on the neighboring sites, the models are called context dependent models. It is well-known that the substitution of nucleotides does not occur independently of neighboring nucleotides, e.g. the CpG effect where an excess of substitutions is observed at positions with a CpG dinucleotide.

Jensen and Pedersen(2000) described the context dependent model at the codon sequence, where the rate of substitution at a site depends on the states at neighboring sites.

They determined the stationary distribution of the Markov process is a Gibbs measure and developed an MCMC method for estimating the transition probability between sequences under the model. Pedersen and Jensen(2001) suggested that in some parts of the genome one sees less *C*s followed by a *G* than expected from the nucleotide frequencies. They discussed the relation between reversibility and the Markov property of the stationary measure. They also proposed a Markov chain Monte Carlo (MCMC) method to evaluate likelihood ratios in the case of two sequences. It is a fairly slow procedure making it less feasible for multiple comparison of sequences. Huttley(2004) incorporated dinucleotide effects into codon substitution models. He considered CpG effects in both transition and transversion substitution rates. For $q_{ij}(i \neq j)$, he proposed the following transition codon matrix.

$$q_{ij} = \begin{cases} 0, & \text{more than one change} \\ \pi_j, & \text{synonymous transversion} \\ \pi_j G, & \text{synonymous transversion involving CpG} \\ \pi_j K, & \text{synonymous transition} \\ \pi_j KG, & \text{synonymous transition involving CpG} \\ \pi_j R, & \text{nonsynonymous transversion} \\ \pi_j RG, & \text{nonsynonymous transversion involving CpG} \\ \pi_j KR, & \text{nonsynonymous transition} \\ \pi_j KRG, & \text{nonsynonymous transition involving CpG.} \end{cases}$$

where the *G* is the *CpG* substitution rate and other notations consistent with Yang's(1998) model.

Christensen et al. (2005) proposed the context codon model. In their model, they not only considered the CpG effect, but also considered how each nucleotide within a codon depends on the two flanking nucleotides.

If a codon sequence x with n codons is written as $x = (x_1, \ldots, x_n)$. To address the three nucleotides of codon x_k, we write $x_k = (x_k^1, x_k^2, x_k^3)$. When nucleotide x_k^j is replaced by another nucleotide, we write the resulting codon as $\tilde{x}_k$.

The rate γ_j for a substitution of nucleotide x_k^j by z depends on the two codons x_k and $\tilde{x}_k$ as well as the two flanking nucleotides of x_k^j. Then the substitution rate for codon x_j is defined as:

$$\gamma_j = Q(x, \tilde{x}) R_j(x_k^j, z; x_k^{j-1}, x_k^{j+1}) \tag{1.12}$$

$$\tag{1.13}$$

where Q specifies a site independent codon model without CpG effect, and R_j relates to the CpG effect. They derived a pseudo-likelihood for the codon substitution models and constructed a corresponding EM-algorithm. They considered a codon model mainly for the analysis of two species. The context dependency is through a CG depression across codon boundaries.

Under the pseudo likelihood approval, the contribution from the ith codon is calculated as though the evolutionary history of the two flanking nucleotides is known. The true evolutionary history for a flanking nucleotide is approximated by either a history with no changes (if the nucleotides in the two sequences are identical) or a history with one change in the middle of the time interval (if the nucleotides in the two sequences are different). Christensen et al.(2005)made a comparison with the full analysis and demonstrated that estimates obtained from the pseudo likelihood are approved very close to the maximum likelihood estimates.

1.5 Aim and organization of the thesis

When context is taken into consideration, the number of independent parameters in substitution models increases dramatically. This makes the estimation of the substitution rate computationally expensive. To understand the effect of context in the substitution models in DNA evolution, more research on this topic is needed. We focus our work on the following aspects:

(1) When dealing with a large number of substitution matrices, most existing work attempt to reduce the number of independent parameters in the same manner for all the matrices. Normally, constraints are added to the rate matrices, such as reversibility and strand symmetry. These are crude approximations, since the true matrices need not obey such constraints. In our work, we adopt an alternative approach. Instead of reducing the number of parameters in each matrix, we reduce the total number of parameters by reducing the number of context dependent matrices.

(2) Parsimony is frequently used in estimation of independent substitution models. We shall adopt the same approach in the estimation of context dependent models, with

a view on improving its performance.

(3) In the estimation of substitution rate matrices, previous work involving the maximum likelihood approach used the EM algorithm and Bayesian MCMC. These methods can be very slow. We intend to use direct optimization of the pseudo-likelihood approach with a view on improving the speed.

(4) Previous research utilized limited real data. In our work, we shall develop a process to simulate context dependent DNA sequence evolutions. This provides the flexibility for doing various experiment using simulated data.

We evaluate the performance of context dependent model via a comparative approach. The present work emphasizes the role of simulation in investigateing the context dependent substitution problem.

In Chapter 2, we introduce the independent substitution process and describe the general context dependent model. We investigate a special case, the two-flanking sites context dependent model. We also propose methods to obtain two specific submodels by reducing the number of matrices.

In Chapter 3, we introduce estimation and evaluation methods. First, we describe two estimation methods: parsimony and maximum pseudo-likelihoods, using the Newton method to maximize the pseudo-likelihood. Then we cover the simulation process. Finally, we describe evaluation methods.

In Chapter 4, we focus on our experiments for simulation data set. We conduct simulation to test the performance of the pseudo-likelihood method against the parsimony method. We then test the parsimony method for context dependent model, and assess model adequacy by goodness-of-fit tests.

In Chapter 5, we apply the pseudo-likelihood method to some real data. We use a clustering method to reduce the number of matrices. We then conduct goodness-of-fit tests for our context dependent models and compare the performance of different models.

We conclude the present work in Chapter 6 and provide some possible directions of further research.

Chapter 2

The general context dependent substitution model

In this chapter, we first use the continuous Markov process to describe the substitution process in a DNA sequence. We then propose our context dependent substitution model.

2.1 Substitution process

Let us consider a DNA sequence of length n, and assume that only one nucleotide changes at a time in the evolution of that sequence. At each site in the sequence, nucleotide substitution is assumed to follow a continuous time Markov process.

Mathematically, we denote the evolution process of a sequence as $\{X(t) : t \geq 0\}$, where t is the evolution time. Thus, $X(t) = X_1(t)X_2(t)...X_n(t)$ is a random sequence and $X_i(t)$ is the base at position i at time t.

2.1.1 Independent substitution process

Before discussing the general case of substitution process, let us first look at the independent substitution process. Assuming the each site evolves is independently during DNA substitution process, here we consider sequence length of $n = 1$ and $X(t)$ is a base.

The continuous Markov process works as follows. The nucleotide at a site stays in one particular state for some time; then at a substitution happens and the nucleotide changes to another. The evolution process is based on a 4×4 substitution rate matrix Q. The waiting time τ for a state change at a site follows an exponential distribution:

$$\tau \sim \exp(-Q(X(t), X(t))). \tag{2.1}$$

For example, if $X(t) = T$, then rates for a substitution from T to C, A, and G are $Q(T, C)$,$Q(T, A)$ and $Q(T, G)$ respectively. The waiting times for a substitution of T to the three types (C,A,G) are $\tau_{T,C}$, $\tau_{T,A}$, $\tau_{T,G}$ respectively. The latter are assumed to be exponentially distributed as follows:

$$\tau_{T,C} \sim \exp(-Q(T, C));$$

$$\tau_{T,A} \sim \exp(-Q(T, A));$$

$$\tau_{T,G} \sim \exp(-Q(T, G)).$$

$\tau_{T,C}$, $\tau_{T,A}$, $\tau_{T,G}$ are independent and the time τ_T follows an exponential distribution,

$$\tau_T \sim \exp(-Q(T,T)) \tag{2.2}$$

where $Q(T,T) = -(Q(T,C) + Q(T,A) + Q(T,G))$. Then τ_T, the waiting time for a substitution of T to any other type, is the minimum of $\tau_{T,C}$, $\tau_{T,A}$ and $\tau_{T,G}$.

2.1.2 General context dependent substitution process

We now describe a general context dependent substitution process on a DNA sequence of length n. In the general substitution process, the state space of the process is $S = \{A, C, G, T\}^n$. We assume that the generic state is $s = (s_1, s_2, \ldots, s_n)$, and that only one site change at a time is permitted during the evolution process. The parameter $Q(x, y; s_{-i})$ is defined as the rate when the state changes from x to y at site i in context s_{-i}, where $x \neq y$, $s_{-i} = s$ with s_i unspecified, that is, $s_{-i} = (s_1, \ldots, s_{i-1}, *, s_{i+1}, \ldots, s_n)$, $i = 1, \ldots, n$.

If we fix $1 \leq i \leq n$, $s_{-i} = (s_1, \ldots, s_{i-1}, *, s_{i+1}, \ldots, s_n)$, let $x = s_i$ and $y \neq x$, then $s = (s_1, s_2, \ldots, s_{i-1}, x, s_{i+1}, \ldots, s_n)$ and $s' = (s_1, s_2, \ldots, s_{i-1}, y, s_{i+1}, \ldots, s_n)$. Our general context dependent substitution model is defined as follows:

$$Q(s, s') = Q(x, y; s_{-i})$$

.

From the above definition, we know that the substitution matrix depend on x, y and its context at i site.

For each site s_i, the waiting time for a change from state x to state y is $\tau_{x,y}$.The distribution of $\tau_{x,y}$ is exponential,

$$\tau_{x,y} \sim \exp(Q(x, y; s_{-i})) \tag{2.3}$$

The possible changes in a sequence can be represented with a graph, in which there are 4^n nodes. As the state at each of the n sites can change to any other three states, each node has $3n$ neighbors in the graph. The substitution process is like a random walk on the graph.

2.2 Special cases

We have described the general context dependent model. The context dependent models proposed by other works can be considered special cases of the general model.

2.2.1 Two flanking site model

There are different ways to define context dependence. Christensen (2006) assumed that substitution matrix should satisfy the two conditions: (1) The substitution matrix depends on immediate neighboring sites only. (2) The substitution matrix is position-invariant, i.e. the substitution matrix does not depend on the position of a site in the sequence. Their model is given by,

$$Q(x, y; s_{-i}) = Q(x, y; s_{i-1}, s_{i+1}) \tag{2.4}$$

Their model is a special case of model (2.3) when the context consists of just the immediate neighbouring bases, instead of the whole sequence. The model (2.4) has $4 \times 4 \times 4 \times 3 = 192$ parameters. In their model, for a sequence with length n, s_0 and s_{n+1} are not defined. In order to allow for s_1 and s_n to change, we define s_0 and s_{n+1} and assign a fixed value A for the two undefined sites, i.e. $s_0 = A$ and $s_{n+1} = A$.

Christensen (2006) model is also called the two-flanking site dependent model because it only depends on the left and right sites. In our work, we will use this model as an example. Our simulation and experiments will be based on this model.

2.2.2 Dinucleotide model

The overlapping dinucleotide model proposed in Lunter and Hein (2004), which allows only single nucleotide substitutions, is a special case of model (2.4) where

$$Q(x, y; s_{-i}) = Q^{left}(x, y, s_{i-1}) + Q^{right}(x, y, s_{i+1}) \tag{2.5}$$

where $Q^{left}(x, y; s_{i-1})$ is the rate of y substituting x when the left neighbor is s_{i-1}, and $Q^{right}(x, y; s_{i+1})$ is the rate of y substituting x when the right neighbor is s_{i+1}.

There are 84 free parameters in this model. In the sequence $s = (s_1, s_2, \ldots, s_n)$, when considering the change of s_i, we use s_{i-1} and s_{i+1} to choose a substitution matrix. Since both s_{i-1} and s_{i+1} have four choices (A, C, G, T), there are 16 combinations of s_{i-1} and s_{i+1} to represent the context of s_i. For a triplet (a, b, c), if we consider the substitution of b, we look for the substitution matrix $Q_{a,c}$. Totally, we define 16 context dependent Q matrices.

The model of 16 context dependent substitution rate matrices is considered as the full model and is referred to as the $16Q$ model hereafter.

2.2.3 Independent model

Model (2.3) also applies to site-independent models. In site-independent substitution models, no context is considered. Therefore, the substitution rate is simplified as $Q(x, y; s_{-i}) = Q(x, y)$. As x has four choices (A,C,G,T) and y ($y \neq x$) has three choices, 12 substitution rates are enough for describing the substitution process. In general, let $X(0) = s = s_1 \dots s_n$; there are $3n$ independent waiting times.

2.3 Clustering of rate matrices

In the discussion of two-flanking site context dependent substitution, for a given site, we consider the two neighboring sites (left and right sites) as the major context. For example, in the sequence $s = (s_1, s_2, \dots, s_n)$, when we look at the change of s_i, we will also look at s_{i-1} and s_{i+1}. For different context (s_{i-1} and s_{i+1}), the substitution rate of s_i are likely to be different. Therefore, we define different substitution matrices for different contexts.

To consider the immediate context, we have to define 16 context dependent Q matrices. To build context dependent substitution models, we need to estimate 16 4×4 substitution rate matrices. Since in each substitution matrices, there are 12 independent rates, we need to estimate $12 \times 16 = 192$ parameters, a very complicated and computationally difficult task. Therefore, we must reduce the number of parameters to reduce the amount of computation effort.

Previous work on context dependent models tried to reduce the number of parameters in each substitution matrix. In this work, we will take an alternative approach. We will reduce the number of matrices instead of the number of parameters in each matrix.

Among the 16 substitution matrices, some matrices may have similar values. To simplify the estimation process and reduce the number of parameters to be estimated, we propose merging some context dependence cases. There are two ways to do so. One way is to use existing knowledge on how DNA substitution happens, e.g. the CpG effects in DNA substitution. Another way is to use statistical approaches to cluster the rate matrices based on the similarities between pair of matrices.

2.3.1 Grouping to four Q matrices

Since the CpG effect is confirmed by previous research (Karlin and Burge, 1995), first we merge the 16 rate matrices into 4 rate matrices: define them by Q_C , Q_G ,Q_{CG} and Q_{others}. The rate matrix Q_{CG} is the rate matrix for the sites, whose left neighboring site is C and right neighboring site is G; the matrix Q_C is the rate matrix for the sites, whose left neighboring site is C and right neighboring site is not G; the matrix Q_G is the rate matrix for the sites, whose right neighboring site is G and left neighboring site is not C; the matrix Q_{others} is the rate matrix for the rest cases. Altogether there are $12 \times 4 = 48$ parameters to estimate in the four Q matrices.

The four-matrix model is referred to as the $4Q$ submodel hereafter.

2.3.2 Grouping to two Q matrices

In order to further simplify the model, we may group our 16 matrices into two matrices. We merge Q_C , Q_G and Q_{CG} together and call the resulting matrix Q_{CorG}. The matrix Q_{CorG} is the rate matrix for the sites, whose left neighboring site is C or right neighboring site is G. The matrix Q_{others} is the rate matrix for the other cases.

The two-matrix model is referred to as the $2Q$ submodel hereafter.

2.3.3 Statistical clustering of Q matrices

There are a number of clustering methods (Johnson and Wichern 2002), such as joining (tree clustering), two-way joining (block clustering), and k-Means clustering. Here we choose the tree clustering method to group the matrices. The purpose of the tree clustering algorithm is to join together objects into successively larger clusters, using some measure of distance. The result of this type of clustering is a hierarchical tree. To use the hierarchical tree clustering method, we need a method to measure the distance between two rate matrices. In the 4 × 4 matrices, there are 12 independent rate values. Therefore, each matrix can be represented as a 12-dimensional vector. The distance between two matrices can be measured using the Euclidean distance between their 12 dimensional vectors. The distance between the two vectors $X = (x_1, x_2, \ldots, x_n)$ and $Y = (y_1, y_2, \ldots, y_n)$ is given by

$$D(X, Y) = \sqrt{\sum_{i=1}^{12} (x_i - y_i)^2}, \tag{2.6}$$

We will apply this method to real data at section 5.2.

2.4 Summary

In this chapter, we first described our general context dependent model. Then we described some special cases of the general model. We have discussed how to reduce the model parameters. This is accomplished by merging $16Q$ matrices into $4Q$ or $2Q$ matrices. We proposed to use clustering method reduce the number of matrices.

Chapter 3

Estimation and evaluation methods

If we know the initial DNA sequence and its evolved final sequence, we can estimate the substitution rate matrices from the two sequences. This chapter covers the estimation methods for the context dependent substitution rate matrices. First, we propose the modified Parsimony method and modified Pseudo-likelihood method to estimate the substitution rates. The we propose a context dependent simulation algorithm. Finally, we describe the evaluation criteria for estimation methods and substitution models.

3.1 Estimation methods

In this section, we describe the two estimation methods for substitution matrices.

3.1.1 The parsimony method

Camin and Sokal (1965) first introduced the simplest parsimony method. Farris (1970) derived the algorithms for counting changes in parsimony method. In the process of DNA sequence evolution, some sites may change many times and reach to the final state. Parsimony method however ignores the intermediate substitutions and considers the multi-step change as one single change from the initial state to the final state. Therefore, its basic assumption is the minimal substitutions during the evolution from one sequence to another. In our context dependent substitution models, when the number of Q matrices has been determined, the next step is to estimate the matrices from data. The parsimony method has been used to estimate substitution rate in site independent models. In our work, we will use it to estimate context dependent substitution models.

Suppose $V = (v_1 v_2 \ldots v_n)$ is the ancestral sequence of $X = (x_1 x_2 \ldots x_n)$, the count of context dependent substitution is given by

$$C_{l,r}(a, b) = \sum_{i=2}^{n-1} M_{v_{i-1},v_{i+1}}(v_i, x_i). \tag{3.1}$$

where l, r are the left and right neighboring site, respectively; a and b are the state before and after substitution, respectively. Note that $l, r, a, b \in \{1, 2, 3, 4\}$, with the numbers corresponding to A, C, G and T respectively.

$$M_{v_{i-1},v_{i+1}}(v_i, x_i) = \begin{cases} 1, & \text{if } v_i = a, x_i = b; v_{i-1} = l, v_{i+1} = r; \\ 0, & \text{others} \end{cases} \tag{3.2}$$

The parsimony method calculates the substitution rate matrices by counting the number of substitutions that have taken place for the context. However, when considering the substitution of a site, its context may also change. This makes substitution counting tricky. There may be different schemes for counting. In our work, we propose the following modified scheme,

$$C_{l,r}(a,b) = \sum_{i=2}^{n-1} (M_{v_{i-1},v_{i+1}}(v_i, x_i) + M_{x_{i-1},x_{i+1}}(v_i, x_i)), \tag{3.3}$$

where

$$M_{v_{i-1},v_{i+1}}(v_i, x_i) = \begin{cases} 1/2, & \text{if } v_i = a, x_i = b; v_{i-1} = l, v_{i+1} = r; \\ 0, & \text{others.} \end{cases} \tag{3.4}$$

$$M_{x_{i-1},x_{i+1}}(v_i, x_i) = \begin{cases} 1/2, & \text{if } v_i = a, x_i = b; x_{i-1} = l, x_{i+1} = r; \\ 0, & \text{others.} \end{cases} \tag{3.5}$$

The transition probability is given by

$$P_{l,r}(a,b) = \frac{C_{l,r}(a,b)}{\sum\limits_{x=1}^{4} C_{lr}(v,x)}. \tag{3.6}$$

When we group the rate matrices into four matrices: Q_C, Q_G, Q_{CG} and Q_{others}, the transition matrices are calculated as follows.

$$P_C(a,b) = \frac{\sum\limits_{r=1}^{4} C_{2,r}(a,b)}{\sum\limits_{r=1}^{4} \sum\limits_{x=1}^{4} C_{2,r}(v,x)}; \tag{3.7}$$

$$P_G(a,b) = \frac{\sum\limits_{l=1}^{4} C_{l,3}(a,b)}{\sum\limits_{l=1}^{4} \sum\limits_{x=1}^{4} C_{l,3}(v,x)}; \tag{3.8}$$

$$P_{CG}(a,b) = \frac{\sum\limits_{l=2 \cup r=3} C_{l,r}(a,b)}{\sum\limits_{l=2 \cup r=3} \sum\limits_{x=1}^{4} C_{l,r}(v,x)}; \tag{3.9}$$

$$P_O(a,b) = \frac{\sum\limits_{l\neq 2\cap r\neq 3} C_{l,r}(a,b)}{\sum\limits_{l\neq 2\cap r\neq 3}\sum\limits_{x=1}^{4} C_{l,r}(v,x)}. \tag{3.10}$$

Subsequently, we can derive the rate matrices based on the transition matrices. For a matrix P, $\log(P)$ is a matrix Q if $\exp(Q) = P$.

Suppose that P is diagonalizable: $P = VD(\lambda_i)V^{-1}$, where V is a 4×4 matrix and $D(\lambda_i)$ is diagonal matrix with the positive eigenvalues $\lambda_1 \ldots \lambda_n$ down the diagonal. Now, generally $Q = \log(P) = V \log D(\lambda_i)V^{-1}$. Thus we can write

$$Q = V \log D(\lambda_i)V^{-1} = V \begin{vmatrix} \log\lambda_1 & 0 & \ldots & 0 & 0 \\ 0 & \log\lambda_2 & \ldots & 0 & 0 \\ \vdots & \vdots & & \vdots & \vdots \\ 0 & 0 & \ldots & 0 & \log\lambda_n \end{vmatrix} V^{-1}$$

Diagonalizing P is not always possible. However in our work, we have not had much problem due to the nature of transition matrix. In DNA sequence substitution process, substitution happens on a small fraction of sites. Therefore, the diagonal elements of transition matrix, which mean the probabilities of unchanged sites, are always larger than other elements in the same row. Subsequently, it is unlikely that the eigenvalues are less than zero. In our real data, we did not find any case that an eigenvalue is non-positive.

The substitution rate matrices are as follows.

$$Q_C = \log P_C;$$

$$Q_G = \log P_G;$$

$$Q_{CG} = \log P_{CG};$$

$$Q_{others} = \log P_{others}.$$

3.1.2 The pseudo-likelihood method

Although the parsimony method is a simple method for estimating the substitution matrices, it remains a simple approximation that overlooks the intermediate substitution process. In DNA evolution, some sites may have changed many times before ending at a final sequence. The likelihood approach provides a means for addressing this issue.

We will use pseudo-likelihood in our work. A pseudo-likelihood function was introduced by Besag(1975) in the context of a random field. It was defined as a product of conditional likelihoods, each term representing the conditional likelihood for the observation at a particular site given the observations at neighboring sites. Christensen(2006) also derived a similar pseudo-likelihood. We use Christensen's definition in our work.

Let us consider the evolution of the sequence $X = (x_1, \ldots, x_n)$ from the ancestral sequence $V = (v_1, \ldots, v_n)$. Since divergence times and substitution rates cannot be distinguished, the substitution rates are standardized such that evolution happens from time $t = 0$ to time $t = 1$.

If the nucleotides at the flanking positions are l and r, the rate matrix for a single nucleotide position is given by the 4×4 rate matrix

$$Q^{lr}(a, b) = Q(a, b; l, r). \tag{3.11}$$

There are $16\times 12 = 192$ parameters for $\left\{Q^{lr}(a,b); a \neq b, l, r \in (A,C,G,T)\right\}$ in the model. Its diagonal is

$$Q^{lr}(a,a) = -\sum_{b\neq a,} Q^{lr}(a,b). \tag{3.12}$$

The above definition means that, when a changes to b, the Q matrix depends on l and r. If l and r are fixed during the process, it is easy to determine Q. What if l and r changed during the process? Considering the kth nucleotide. If $v_{k-1} \neq x_{k-1}$, there is a change to the left nucleotide, and we assume it happens at time $t = 1/2$. Similarly so for the nucleotide to the right. The substitution matrix for nucleotide k can be approximated as $Q^{v_{k-1},v_{k+1}}$ for $(0 \leq t \leq 1/2)$ and $Q^{x_{k-1},x_{k+1}}$ for $(1/2 \leq t \leq 1)$. That is,

$$Q^{*}(a,b,t) = \begin{cases} Q^{v_{k-1},v_{k+1}}(a,b), & \text{if } 0 \leq t \leq 1/2; \\ Q^{x_{k-1},x_{k+1}}(a,b), & \text{if } 1/2 \leq t \leq 1. \end{cases} \tag{3.13}$$

The likelihood of the observation at position k is defined as

$$L_{v_{k-1},v_{k+1};x_{k-1},x_{k+1}}(x_k \mid v_k) = [\exp(Q^{v_{k-1}v_{k+1}}/2)\exp(Q^{x_{k-1}x_{k+1}}/2)]_{v_k,x_k}. \tag{3.14}$$

The pseudo-likelihood of the observations in the sequence is defined as

$$L = \prod_{k=1}^{n}(L_{v_{k-1},v_{k+1};x_{k-1},x_{k+1}}(x_k \mid v_k)) \tag{3.15}$$

Given the initial sequence and the final sequence, the Q matrices can be obtained by maximizing the pseudo-likelihood function.

3.1.3 Optimization method

To estimate the Q matrices, we need an optimization method to obtain the values of Q matrices that maximize the pseudo likelihood function. In our work, we use the Broyden-Fletcher-Goldfarb-Shanno method (Broyden, 1970) for optimization. The BFGS method is commonly used to solve unconstrained nonlinear optimization problems. It is derived from Newton's method, which is a class of hill-climbing techniques that seek the stationary point of a function. Newton's method assumes that the objective function can be locally approximated as a quadratic Taylor expansion in the region around the optimum. It uses the first and second derivatives to find the stationary point.

In quasi-Newton methods, the Hessian matrix of second derivatives of the function to be minimized does not need to be computed at any stage. Instead, it is updated by analyzing successive gradient vectors. Quasi-Newton methods are a generalization of the secant method, which seeks the roots of the first derivative for multidimensional optimization problems. In multi dimensions the secant equation is under determined, and quasi-Newton methods differ in how they constrain the solution. The BFGS method is one of the most popular members in this class.

Our optimization problem involves many parameters (24 parameters for $2Q$ model, 48 for $4Q$ model and 192 for $16Q$ model). Therefore, good choices of initial values are extremely important for the optimization to arrive at a convergence point. In this work, we first use parsimony method to estimate the rate values. Although the values are not very accurate, they are good enough for using as initial values for optimization.

3.2 Simulation study

We will first describe how an initial DNA sequence is modeled to evolve into a new sequence given the substitution rate matrices. Then we will describe a simulation process in the experiments in the work.

3.2.1 Simulation process

Given an initial DNA sequence and a set of substitution rate matrices for different context cases, we can simulate the context dependent substitution process. A simulated substitution process can convert an initial sequence (ancestral sequence) into a new sequence (descendent sequence). From the ancestral and descendent sequence pair, we can estimate the substitution rate under different models. Therefore, we can evaluate the performance of different models or different estimation methods by comparing the estimated rate matrices with the actual substitution rate matrices that are used for simulation.

The work flow of the simulation process is as follows. Suppose the DNA sequence is $s_{-i} = (s_1, , s_{i-1}, *, s_{i+1}, , s_n)$ at time $t = 0$. The substitution rates are $Q(x, y; l, r)$, $x, y, l, r \in \{1, 2, 3, 4\}$, representing $\{A, C, G, T\}$ respectively. The simulation will start from time $t = 0$ and end at time $t = 1$. To find out the sequence status at time $t = 1$, we work in the following steps.

(1) Generate initial waiting times: Let $t = 0$. Generate random waiting times $T = \{w_1, w_2, ..., w_n\}$ for each site, where t_i follows the exponential distribution:

$$t_i \sim \exp(Q(s_i, s_i; s_{i-1}, s_{i+1})), 1 \leq i \leq n \tag{3.16}$$

(2) Find the site that changes the earliest: If u represents a site that will change, then the earliest change occurs at time $\tau = w_u$, where

$$u = \arg \min_{1 \leq i \leq n} w_i; \tag{3.17}$$

(3) Generate a new state: Generate a new random state y for site u, according to the probability

$$P(y) = -Q(s_u, y; s_{u-1}, s_{u+1})/Q(s_u, s_u; s_{u-1}, s_{u+1}); \tag{3.18}$$

Then update s_u with y.

(4) Update waiting time: First, let w_i $(i = 1, ..., n)$ be updated with $w_i - \tau$, and let t be updated with $t + \tau$. Then generate a new waiting time for the site i, $i = u - 1$, u and $u + 1$, which follows the exponential distribution.

(5) Check termination condition: If the total waiting time t is larger than 1, then output the sequence S and stop. Otherwise, the process repeats from (2).

Simulation data can be used to evaluate estimation methods and substitution models. For a more realistic study, we used the rate matrices derived from the actual data with a parsimony approach.

3.2.2 Simulation based on real data

The simulation uses information of a real database (Huttley, G.A., personal communication), which contains 242 sequence alignments. The length of an alignment ranges from a few thousand to one hundred thousand base pairs. Each alignment consists of three sequences, from the three species: human, chimp, and macaque.

The phylogenetic tree of the three species is shown in Figure 3.1. The common ancestor of human ($s1$) and chimp ($s2$) is $s4$, and the common ancestor of macaque ($s3$) and $s4$ is $s5$. Presently, we focus on the evolution from $s4$ to $s1$ and $s2$. In this thesis, the branch connecting $s4$ to $s1$ is referred to as H branch; and the branch connecting $s4$ to $s2$ is referred to as the C branch.

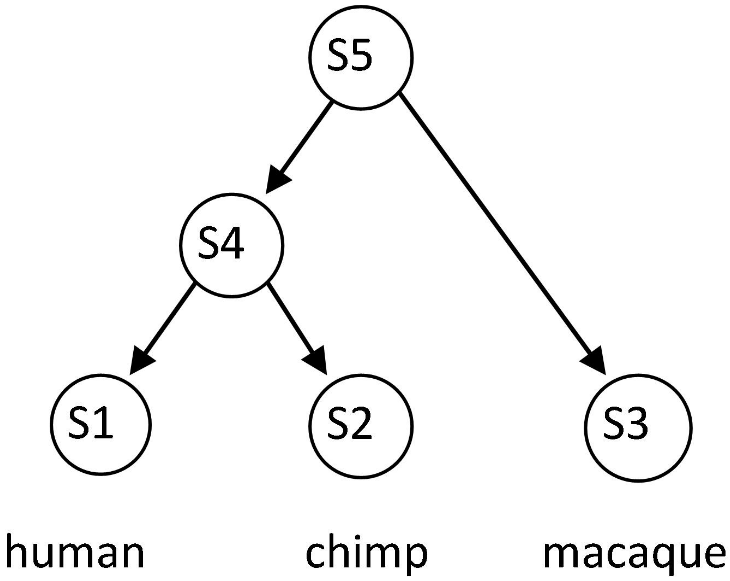

Figure 3.1: Phylogenetic tree for human-chimp-macaque

Our experiment procedure with simulated data is as follows.

Table 3.1: Ancestor inference method

Case	Inferring ancestor	Condition
1	$s_4=s_1$,$s_5=s_1$	if $s_1=s_2$ and $s_1=s_3$
2	$s_4=s_1$,$s_5=s_1$	if $s_1=s_2$ and $s_1 \neq s_3$
3	$s_4=s_1$, $s_5=s_1$	if $s_1 \neq s_2$ and $s_1=s_3$
4	$s_4=s_2$, $s_5=s_2$	if $s_1 \neq s_2$ and $s_2=s_3$
5	$s_4=5$, $s_5=5$	if $s_1 \neq s_2$ and $s_1 \neq s_3$ and $s_2 \neq s_3$
6	$s_4=0$, $s_5=0$	if $s_1=0$ or $s_2=0$ or $s_3=0$

(1) Estimate substitution rate matrices from real data using the parsimony method. For the three sequences in the alignment, we first estimate the common ancestor of human and chimp. We use the conditions in the Table 3.1 to infer the ancestors. This table shows how to infer s4 and s5 from s1, s2 and s3. For example, in case 1, if $s_1=s_2$ and $s_1=s_3$, then we can get $s_4=s_1$ and $s_5=s_1$. Please note that in the table, 0 means gap, 5 means undetermined state. 0 and 5 will be discarded in our calculation.

(2) From the ancestral sequence and the descendent sequence human, we calculate a set of rates (Q_1).

(3) Given a random ancestor sequence of the same length as the real data (V), we simulate the context dependent substitution process with the rate matrices, resulting in a descendent sequence (X) after certain rounds of substitutions.

(4) Estimate the context dependent rates (Q_2) from sequences V and X with the

proposed context dependent models (parsimony or pseudo-likelihood methods).

(5) Calculate the root mean square errors (RMSE) between Q_1 and Q_2. The lower the RMSE, the better the performance of the methods.

3.3 Evaluation methods

3.3.1 Comparing two estimation methods

In statistics,the root mean square error (RMSE) is a frequently-used measure of difference between estimator and the parameter values. It is a simple measure of similarity.

Suppose $X = (x_1, x_2, \ldots, x_n)$ is the estimator vector and $Y = (y_1, y_2, \ldots, y_n)$ is the parameter vector, and both of their lengths are n, then RMSE is calculated as follows:

$$RMSE(X,Y) = \sqrt{\sum_{i=1}^{n} (x_i - y_i)^2 / n} \tag{3.19}$$

In our work, we use RMSE to judge the desirability of different models or methods. If a model or method has smaller RMSE then it is a better approach. We calculate RMSE between actual substitution rate matrices and the predicted substitution rate matrices. As only the off-diagonal elements are independent parameters, we only consider these elements in our calculation.

Suppose our model has N substitution matrices, each of which is 4×4, and $A = \{a^m_{i,j}\}$ and $B = \{b^m_{i,j}\}$ are predicted substitution rate matrices, where $m \in \{1, \ldots, N\}$ means the

mth matrix in the model, and $i, j \in \{1, 2, 3, 4\}$. Then the RMSE between A and B is given by

$$RMSE(A, B) = \sqrt{\sum_{m=1}^{N} \sum_{i=1}^{4} \sum_{j \neq i} (a_{i,j}^{m} - b_{i,j}^{m})^2 / 12N} \tag{3.20}$$

3.3.2 Comparing two models

The likelihood ratio test (LRT) is a statistical test of the goodness-of-fit between two models. Basically, a relatively more complex model is compared to a nested simpler model to see if it fits a particular data set significantly better. As far as possible, a simpler model is to be preferred over a more complex model. Since the more complex model has more parameters, it always returns a higher likelihood score. The LRT requires the log-likelihood score to exceed a certain value before declaring the simple model as inadequate.

The *LR* statistic is given by

$$LR = -2(\log L_2 - \log L_1), \tag{3.21}$$

where $\log L_1$ and $\log L_2$ are the likelihood values for the simple model and complex model, respectively. The distribution of (3.20) statistic approximately follows a chi-square distribution.

To determine if the difference in likelihood scores among the two models is statistically significant, we next must consider the degree of freedom. In the LRT, degrees of freedom is equal to the number of additional parameters in the more complex model. Using this information we can then determine the critical value of the test statistic from standard statistical tables.

The LRT is used to test a simple null hypothesis against a simple alternative hypothesis. Thus, we can use the likelihood ratio test to compare substitution models. In our work, we extend the likelihood ratio test to the pseudo-likelihood ratio test.

3.4 Summary

In this chapter, we have described two estimation methods: parsimony and maximum pseudo-likelihood. We modified the parsimony method to suit context dependent problem by dealing with the changes of context sites. For the pseudo-likelihood method, we uses the BFGS optimization method to estimate the parameters. We have also described the simulation algorithm of context dependent substitution process. We proposed the RMSE method for comparing two methods and the LRT for comparing two models, one nested with the other.

Chapter 4

Numerical study on simulation data

In this chapter, we will perform numerical study through simulation study. We first test the modified parsimony method. Then we compare the two estimation methods: the parsimony method and the pseudo-likelihood method. Finally we compare the two sub-models ($2Q$ model and $4Q$ model) of the two-flanking sites dependent model.

4.1 Numerical simulation for parsimony method

The parsimony method is originally introduced for use in independent substitution models. In the previous chapter, we have proposed a new counting method in the parsimony estimation process to handle the change of context in the evolution process.

We undertook simulation experiments to show the effectiveness of our proposed method. We used the $2Q$ model both for simulation and estimation. The flow of the simulation is as follows:

(1) For each alignment in the real data, we used the simple parsimony method for calculating the substitution rate matrices of $2Q$ model. The estimated substitution matrices were used to generate simulated data.

(2) We multiplied the reference substitution rates by 100 so that more substitutions in the simulated sequence could be observed, allowing us to assess how the change of context affected the estimation accuracy.

(3) We generated a random sequence of length 100000, in which the types of nucleotide followed uniform distribution. Using this random sequence as ancestral sequence, we then generated simulative descendant sequence according to the augmented rates in (2).

(4) Using the $2Q$ model and the parsimony method, we obtained the estimated rate matrices. Two parsimony methods that differ in the way of counting the number of substitutions were both used. The first one is the normal way of counting; the second one is our proposed counting method, in which context changes during evolution.

(5) Finally, we calculated the RMSE values between the off diagonal elements of the simulation rate matrices and those of the estimated rate matrices.

The results of our experiment are shown in Figures 4.1 and 4.2. The normal way of count is labeled as “1 count” , and the proposed method is labeled as “0.5 count”. In the figures, each point represents the RMSE values obtained with the two methods for one alignment. We observed that the most of the RMSE values for “0.5 count” are lower than RMSE values for “1 count” for the same alignment. Therefore, the proposed ”0.5 count” method is significantly better than the normal ”1 count” method.

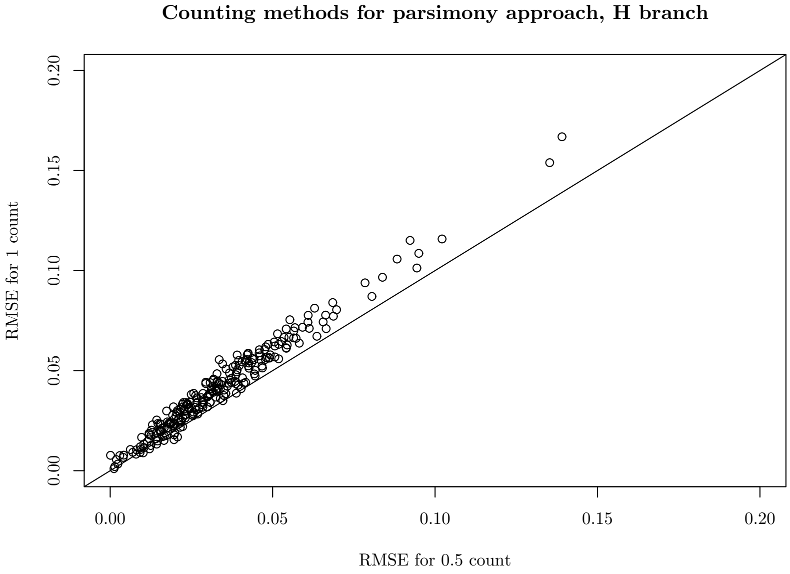

Figure 4.1: Counting method for parsimony approach, H branch

4.2 Comparison of parsimony and maximum pseudo-likelihood methods

We conducted experiments on simulation data generated with real data as reference using both the $2Q$ and $4Q$ models. We examined which estimation method was better between the parsimony and the pseudo-likelihood methods for different models.

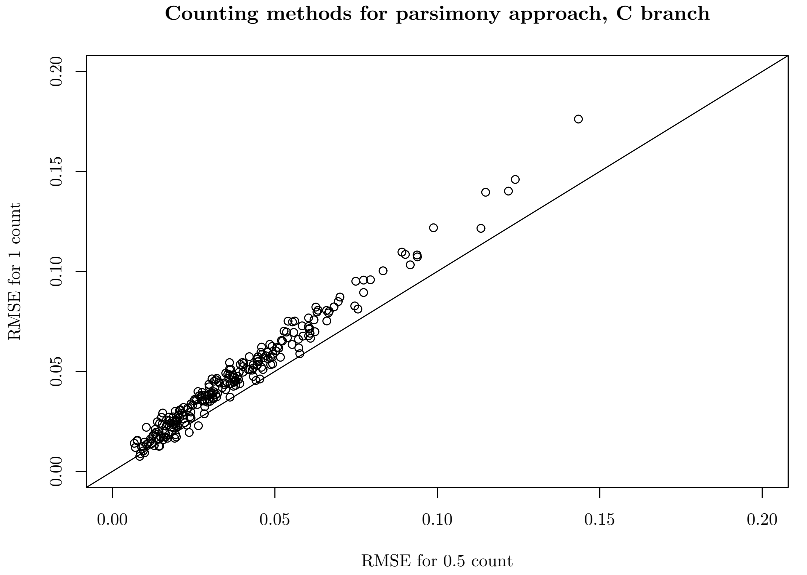

Figure 4.2: Counting method for parsimony approach, C branch

4.2.1 Simulation based on $2Q$ model

We first performed experiments on the $2Q$ model, using substitution rate matrices estimated from real data, and then compared the parsimony and the pseudo-likelihood methods by evaluating their RMSE values.

Some sample RMSE values of the two methods under $2Q$ model are given in Table 4.1. From the table, it can be seen that the two estimation methods yield identical values for both branches under $2Q$ model. The scatter plots for the two branches are as shown in Figure 4.3 and Figure 4.4. In the figures, the coordinates of each point represent the RMSE results obtained with the two methods.

From these two figures, we see that the RMSE values of two methods are identical, implying little, if any, difference between the parsimony and the pseudo-likelihood

Table 4.1: RMSE of parsimony and pseudo-likelihood for 2Q

	H branch		C branch	
Sample	Parsimony	Pseudo-likelihood	Parsimony	Pseudo-likelihood
i	0.000397	0.000397	0.000295	0.000295
ii	0.000638	0.000638	0.000358	0.000358
iii	0.000486	0.000486	0.000385	0.000385
iv	0.000412	0.000412	0.000703	0.000703
v	0.000090	0.000090	0.000408	0.000408

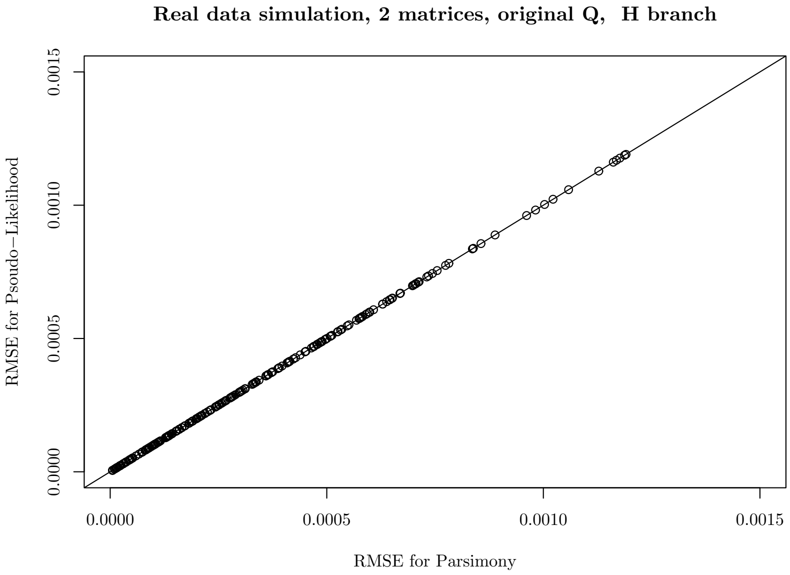

Figure 4.3: RMSE of 2Q simulation for H branch

methods in the the present simulation.

We noticed that the estimated substitution rates Q were very small (the range of 0.001 to 0.01). To find out how the two methods perform in a longer term or with higher

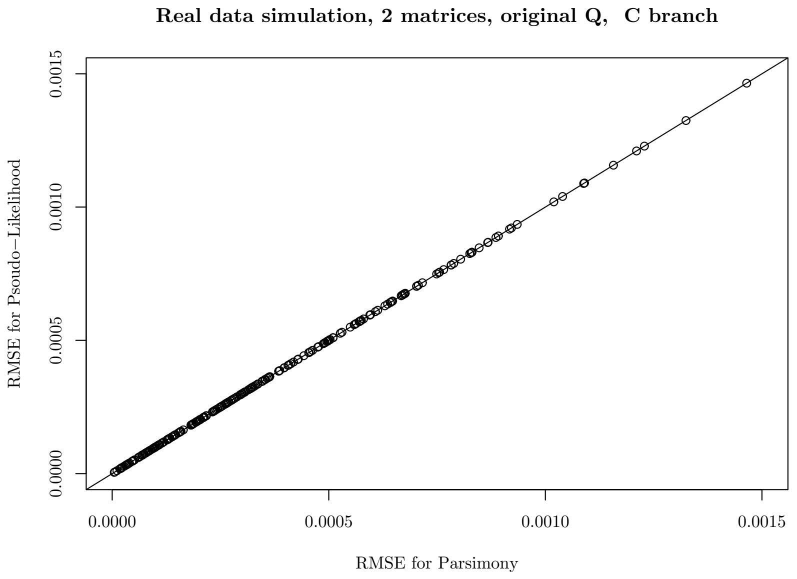

Figure 4.4: RMSE of 2Q simulation for C branch

substitution rates, we increased our substitution rates to 100 times and then repeated the same experiments. That means, if Q is the rate matrix estimated from the real data, we used $Q\times100$ instead of Q to in the simulation, and all the comparisons were done using $Q\times100$.

Several sample results of this simulation are shown in Table 4.2. From Table 4.2, we can see that the RMSE values of pseudo-likelihood are all less than that of the parsimony for both the H branch and the C branch. This means the pseudo-likelihood method is much better than the parsimony method when we use $Q\times100$ to do simulation under $2Q$ model.

The scatter plots for comparing parsimony and pseudo-likelihood under the two branches using RMSE are shown in Figure 4.5 and Figure 4.6. In the figures, the coor-

Table 4.2: RMSE of parsimony and pseudo-likelihood for 2Q, $Q \times 100$

	H branch		C branch	
Sample	Parsimony	Pseudo-likelihood	Parsimony	Pseudo-likelihood
i	0.104301	0.065873	0.099809	0.049148
ii	0.019737	0.006874	0.032963	0.018975
iii	0.035016	0.021076	0.016474	0.002678
iv	0.061378	0.036994	0.033951	0.013645
v	0.040059	0.018199	0.070878	0.042003

dinates of each point represent the RMSE results obtained with the two methods.

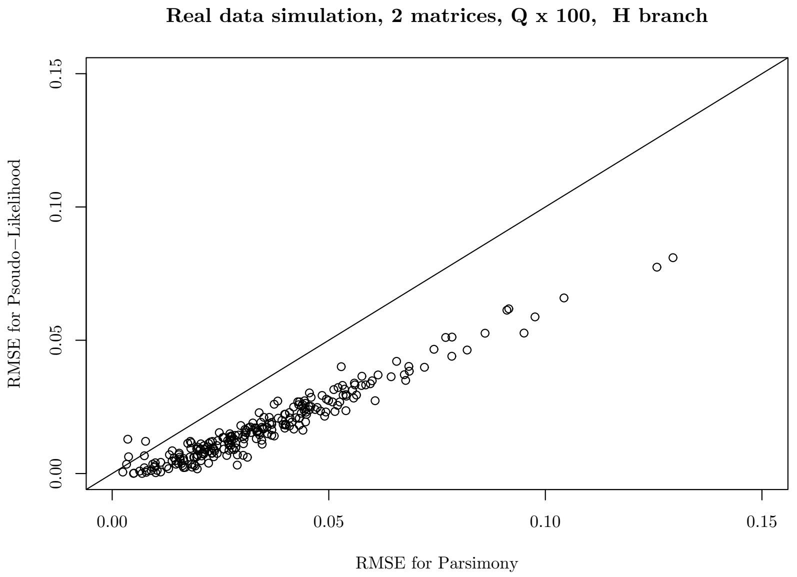

Figure 4.5: RMSE of 2Q simulation for H branch ($Q \times 100$)

From Figures 4.5 and 4.6, we see that the RMSE of pseudo-likelihood are smaller than that of the parsimony method, implying that the pseudo-likelihood method is much

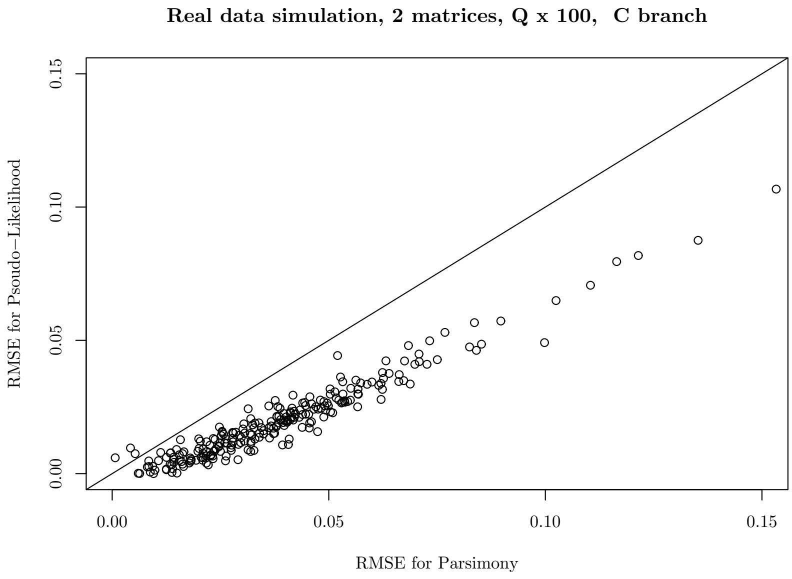

Figure 4.6: RMSE of 2Q simulation for C branch ($Q \times 100$)

better than the parsimony method.

It is known that in independent substitution models, if the substitution rate is small over evolutionary time, the parsimony method is justified (Durbin et. al. 1998, 173-179). However, if the substitution rates are moderate or large, the parsimony method may fail. That implies that when we increase the substitution rate, the parsimony method should be worse than likelihood method. Our simulation results permit us to draw similar conclusion in context dependent substitution models.

4.2.2 Simulation based on $4Q$ model

We also did simulation under $4Q$ model. Some sample RMSE values for comparing parsimony and pseudo-likelihood under $4Q$ model are given in Table 4.3 for the two

Table 4.3: RMSE of parsimony and pseudo-likelihood for 4Q

	H branch		C branch	
Sample	Parsimony	Pseudo-likelihood	Parsimony	Pseudo-likelihood
i	0.013568	0.008948	0.000669	0.000040
ii	0.007930	0.003736	0.002566	0.000753
iii	0.023235	0.008940	0.008741	0.001003
iv	0.011128	0.006217	0.001235	0.000349
v	0.017736	0.006237	0.025901	0.000543

branches. We see that the RMSE value of pseudo-likelihood is smaller than that of the parsimony, implying superiority under $4Q$ model for the two branches.

Scatter plots (Figure 4.7, 4.8) for the two branches show the difference between the parsimony and the pseudo-likelihood under $4Q$ model. In the figures, the coordinates of each point represent the RMSE results obtained with the two methods.

We conclude from our comparative study that even for small substitution rate Q (range from 0.001 to 0.01), the pseudo-likelihood method works better than parsimony under $4Q$ model.

4.3 Biases of estimation

Up to this point, we have examined the RMSEs for the two methods. Here, we look at the bias of individual substitution rates for each method, which tells us something about

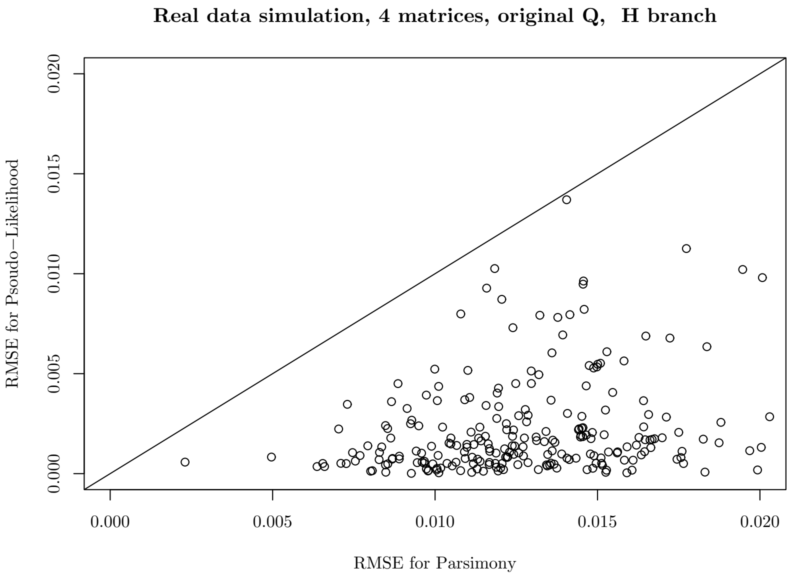

Figure 4.7: RMSE of 4Q simulation for H branch

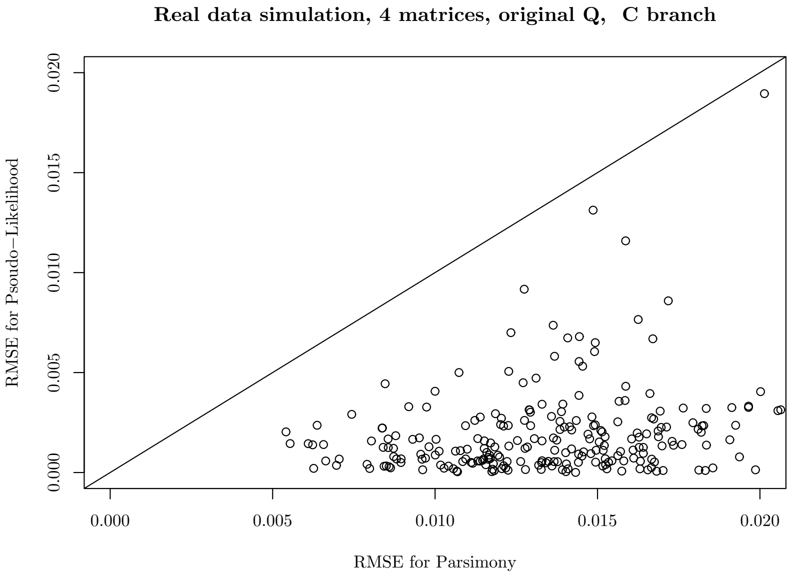

Figure 4.8: RMSE of 4Q simulation for C branch

the accuracy of prediction.

Table 4.4: Index of individual rate in rate vector

	A	G	C	T
A	–	1	2	3
G	4	–	5	6
C	7	8	–	9
T	10	11	12	–

4.3.1 Biases of estimation based on $2Q$ model

As we have two matrices in the sub-model and each matrix has 12 independent rate values excluding the diagonal rates, there are 24 independent rates for two matrices. The 12 rate values of each matrix form a vector, where the rate values in each matrix are organized in the order as indexed in Table 4.4. When the two vectors are joined together, the rate values of the first matrix are indexed from 1 to 12, and those of the second matrix are indexed from 13 to 24.

1. Biases of Parsimony method for $2Q$

Figure 4.9 shows the biases of the parsimony method for the $2Q$ model for *H* branch. From the figure, we can see that the medians of all the rates are almost zero. Therefore, the biases of each rate are small.

2. Biases of pseudo-likelihood method for $2Q$

Figure 4.10 shows the biases of the pseudo-likelihood method for the $2Q$ model for

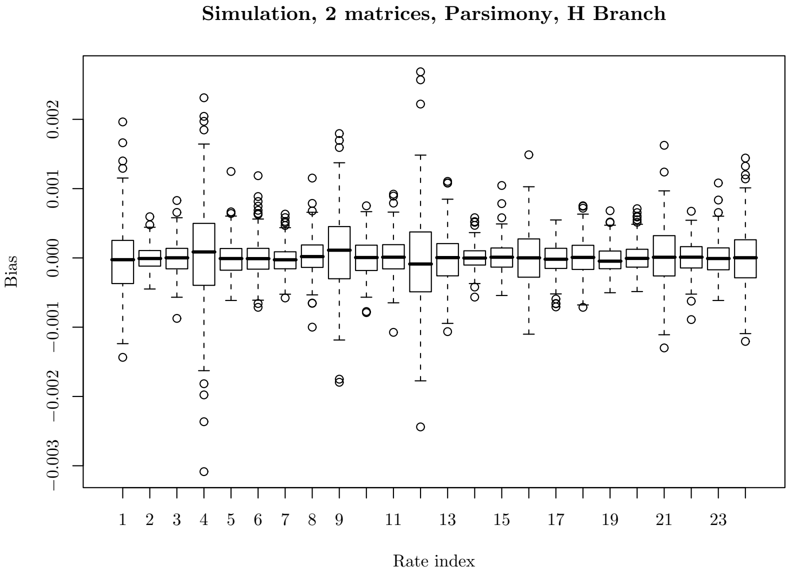

Figure 4.9: Biases of parsimony method for 2Q, H branch

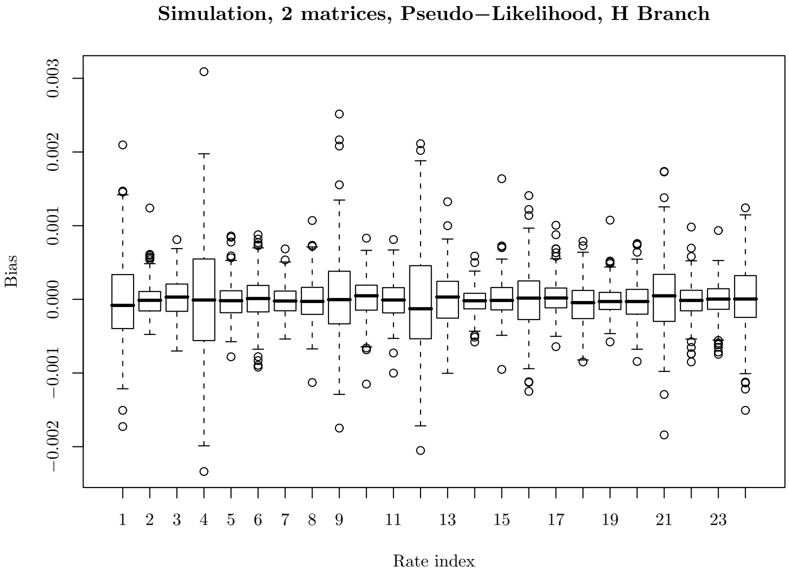

Figure 4.10: Biases of pseudo-likelihood method for $2Q$, H branch

H branch. From the figure, we can see that the medians of all the rates are also almost zero. We also examined *C* branch, and obtained similar results. As the all the biases are small, we can conclude that our two estimation methods under $2Q$ model are accurate.

Although all the biases are small in our two estimation methods, we notice that some rates have a little bigger variance. It seems the transitions of $A \leftrightarrow G$ (e.g. rate index 1, 4, 13, 16) and $C \leftrightarrow T$ (e.g. rate index 9, 12, 21, 24). We examined the rate values for each method and found that these rates actually have larger rate values than others (Refer to Figures 4.11 and 4.12). This shows that in the estimation using our methods, larger rates will have a larger variance.

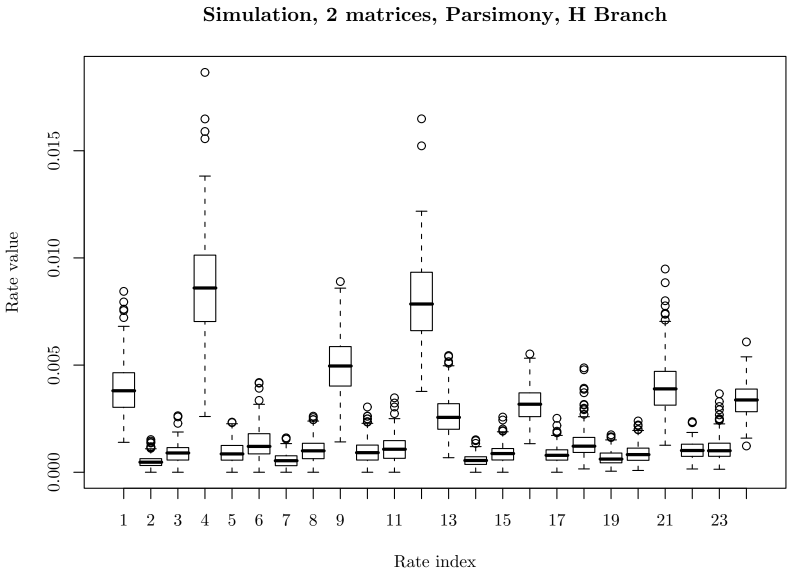

Figure 4.11: Rate values of parsimony method for 2Q, H branch

4.3.2 Biases of estimation based on $4Q$ model

As we have four matrices in the model and each matrix has 12 independent rate values excluding the diagonal rates, there are 48 independent rates for four matrices. The

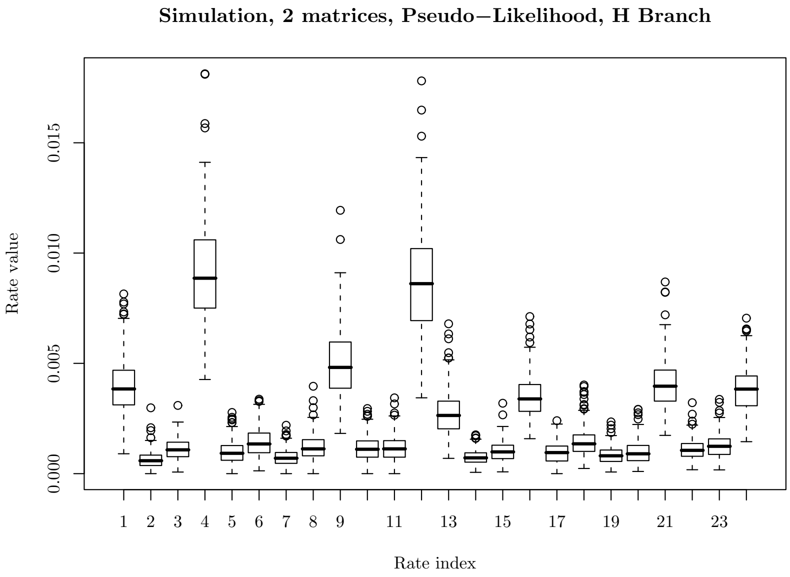

Figure 4.12: Rate values of parsimony method for 2Q, H branch

12 rate values of each matrix form a vector, where the rate values in each matrix are organized in the order as indexed in Table 4.4. Then the four sequences are joined together. In the joint vector, the elements from 1 to 12 represents the first matrix, the elements from 13 to 24 are for the second matrix and so on.

1. Biases of Parsimony method for $4Q$

Figure 4.13 shows the biases of Parsimony method for the $4Q$ model. From the figure, we can see that the medians of all the rates are almost zero. Therefore, all the biases of each rate are small.

2. Biases of pseudo-likelihood method for $4Q$

Figure 4.14 shows the biases of pseudo-likelihood method for the $4Q$ model. From

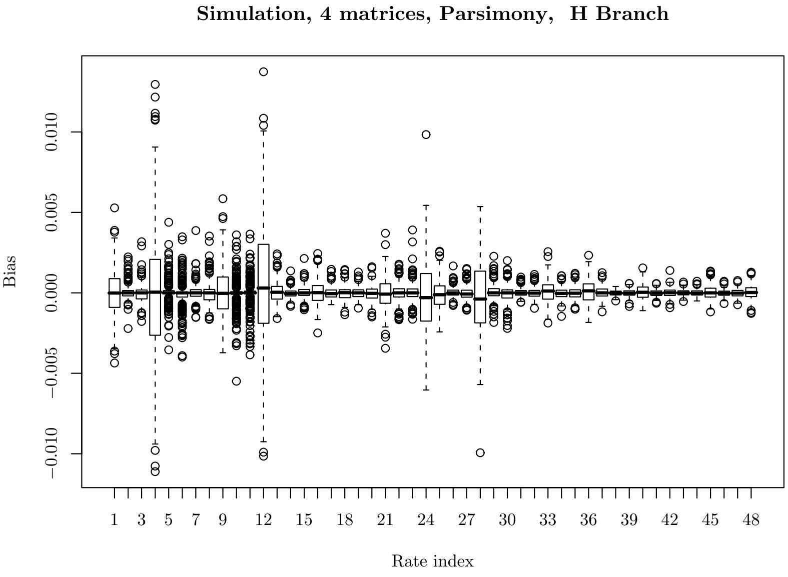

Figure 4.13: Biases of parsimony method for 4Q, H branch

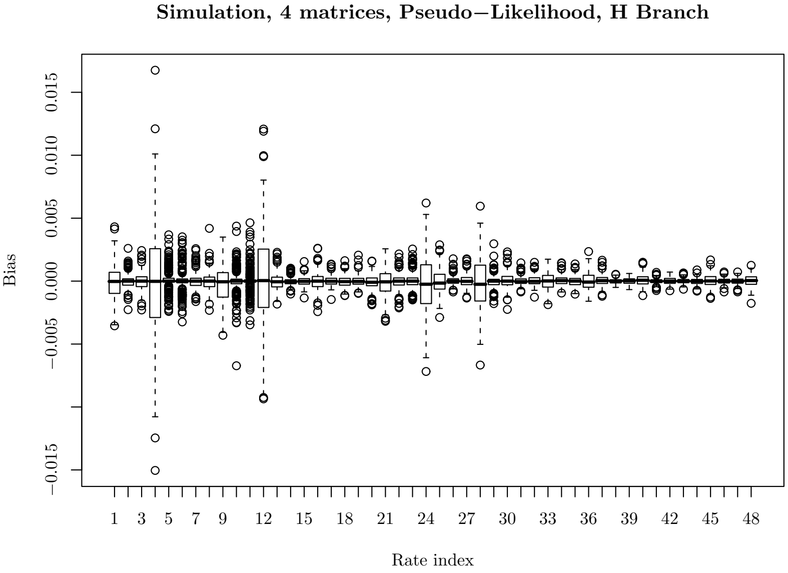

Figure 4.14: Biases of pseudo-likelihood method for 4Q, H branch

the figure, we can see that the medians of all the rates are also almost zero. We also examined the C branch, and obtained similar results. As the biases are small, we can conclude that our two estimation methods under $4Q$ model are accurate.

Similarly, we have examined the rates that have larger variances (e.g. rate index 1, 4, 9, 12, 13, 16, 21, 24, 25, 28) and found that those larger rates normally have a larger variance (refer to Figures 4.15 and 4.16)

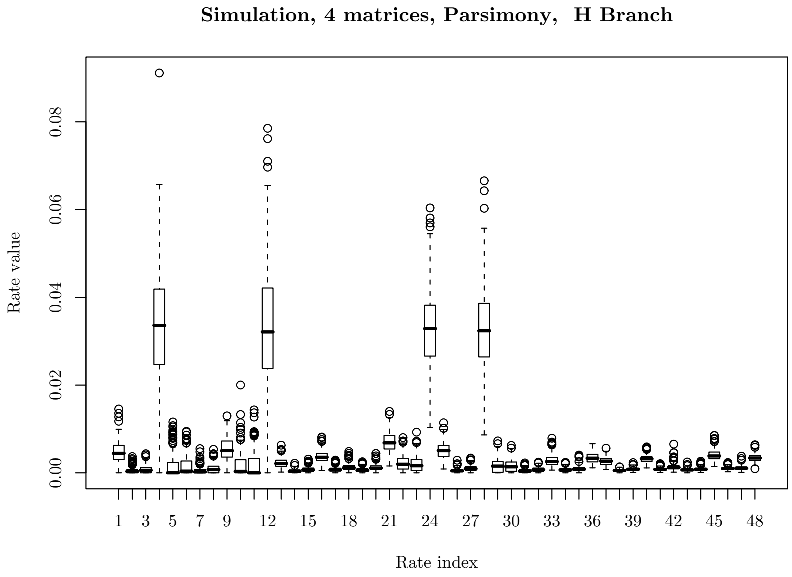

Figure 4.15: Rate values of parsimony method for 4Q, H branch

4.4 Comparison of models

In this section, we use simulation data to test different submodels. We base our simulation on one alignment from real data, with 200 iterations each time. The length of the simulation sequence is chosen as 100000.

In our work, we consider the three hypothesis tests as in Table 4.5. We use H_0,

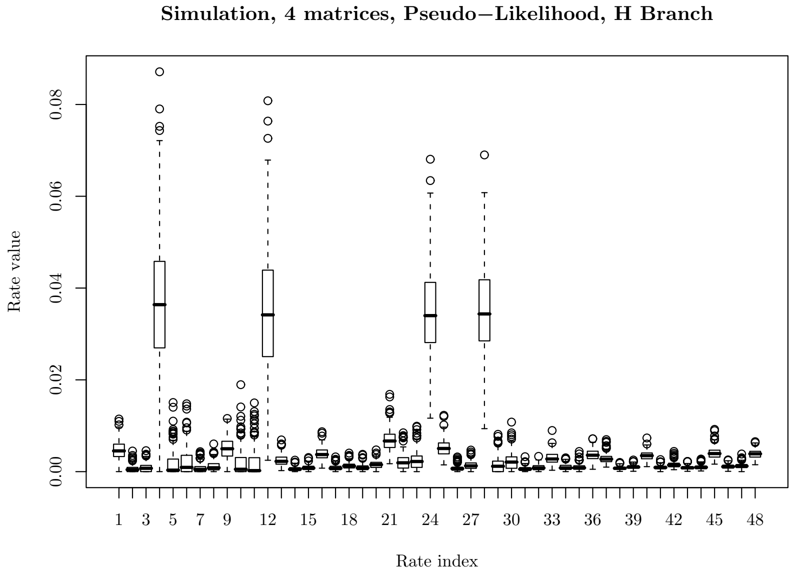

Figure 4.16: Rate values of pseudo-likelihood method for 4Q, H branch

Table 4.5: Hypothesis testing

Test	Null hypothesis	Alternative hypothesis	Degree of freedom
A	2Q model	4Q model	24
B	2Q model	16Q model	168
C	4Q model	16Q model	144

H_1 and H_2 to represent $2Q$, $4Q$ and $16Q$ model respectively. In our test hypothesis, comparison of the two models is directly using the pseudo-likelihood function. We use pseudo-likelihood ratio score instead of likelihood ratio score.

In this section, we only discussed one branch (e.g. H branch) since the results of two branches are very similar.

4.4.1 Approximate distribution

Normally likelihood ratio test is used to compare full models and reduced models. As we use pseudo-likelihood ratio test, we need to first check whether pseudo-likelihood ratio test also follows chi-square distribution asymptotically. For simplicity, in our context, LRT means pseudo-likelihood ratio test.

We can use QQ plot to check whether the distribution of LRT under H_0 follows a chi-square distribution. A QQ plot is a plot of the quantiles of two distributions against each other. It is a graphical method for comparing two probability distributions. In the following QQ plots, the x-axis is quantiles of chi-square distribution, and y-axis is the quantiles of the LRT scores. If the LRT scores follow chi-square distribution, a plot will show a straight line.

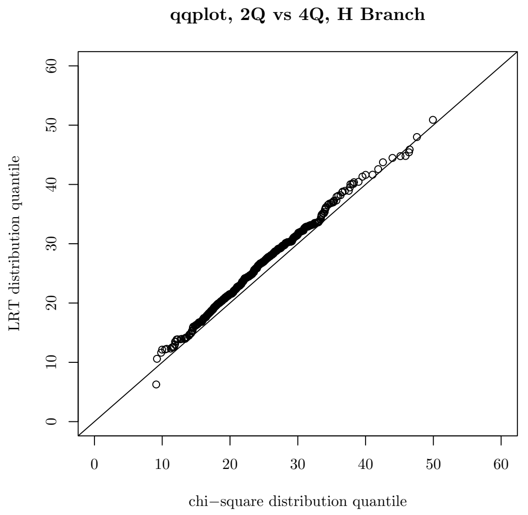

Figure 4.17: Test A: Under 2Q, QQ plot for 2Q vs 4Q, H branch

Figures 4.17, 4.18, 4.19 show the QQ plots for the three hypothesis tests for H

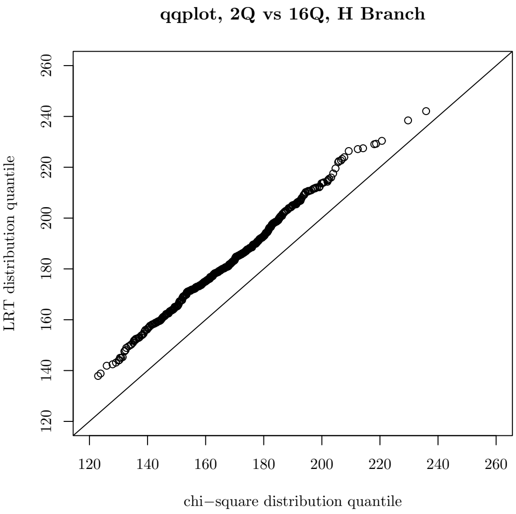

Figure 4.18: Test B: Under 2Q, QQ plot for 2Q vs 16Q, H branch

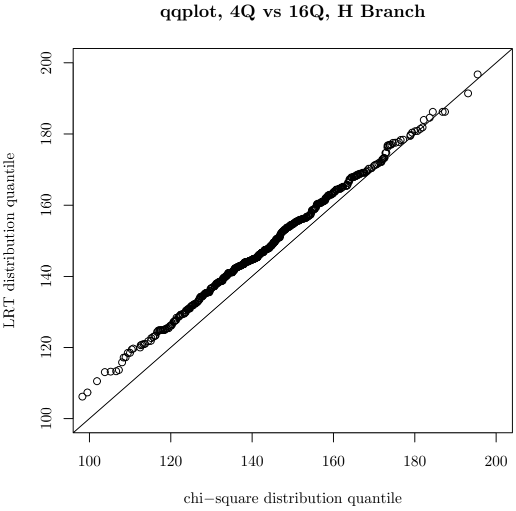

Figure 4.19: Test C: Under 4Q, QQ plot for 4Q vs 16Q, H branch

branch from one simulation and LRT test. From the figures, we can see that each of the figure is almost a straight line for the H branch. But none of the straight lines passes through the origin. This observation suggests that the null hypothesis does not strictly follows a chi-square distribution. However, the LRT score is related to chi-square distribution.

Some previous research proved that an adjusted pseudo-likelihood ratio test follows a chi-square distribution. For example, Geys, Molenberghs and Ryan (1999) proposed the adjusted pseudo-likelihood ratio test statistic LR^* is approximately χ^2(df) distributed, where $LR^* = LR/C$ and C is a constant.

In our work, because we are more interested in the the 95% quantile of the real data, we would examine the 95% quantiles of three hypothesis tests and try to find the new conservative cut-off points in the tests from our simulation results. We randomly selected 20 alignments and repeated the simulation 200 times for each alignment. Each round of simulation is based on one alignment under different models (2Q, 4Q and 16Q). The 95% quantile of LRT scores were calculated. Test results of 8 alignments among the 20 are as shown in Table 4.6. The 8 alignments include those with maximum 95% quantile values.

From the table, we can see that the 95% quantiles of three hypothesis tests are a bit different from expected values (quantiles of chi-square distributions). To be conservative, we use the maximum values of the quantiles of both branches as our critical values in real data testing. From the table, we can see that the maximum values of Tests A, B and C are 40.5, 221.4 and 191.6 respectively.

Table 4.6: 95% quantiles of three hypothesis tests

	TestA		Test B		Test C	
Sample	H branch	C branch	H branch	C branch	H branch	C branch
1	35.8	37.8	204.5	210.8	185.1	172.1
2	34.8	38.4	199.2	218.6	173.5	177.1
3	39.1	36.4	205.4	218.3	182.6	184.5
4	38.2	39.9	212.5	219.8	176.5	183.1
5	37.1	37.8	210.1	206.1	178.1	183.1
6	37.1	40.5	213.9	216.2	181.9	178.3
7	36.9	38.3	221.4	217.6	191.6	186.9
8	37.8	39.1	217.6	213.4	186.3	190.5
χ^2 quantiles	36.4	36.4	199.2	199.2	173	173
maximum	39.1	40.5	221.4	219.8	191.6	191.6

4.5 Summary

From the simulation studies, we conclude that the pseudo-likelihood method is more general and robust than parsimony method. We see that when substitution rates are large, the pseudo-likelihood approach has obvious advantages over the parsimony methods. We have also analyzed the biases of estimation and identified pseudo LRT cutoff values for the 95% percentile.

Chapter 5

Analysis of real data

5.1 Description of the data set

We obtained a real data set consisting of 242 sequence alignments from Dr. Gavin A. Huttley (The Australian National University). The length of the alignment is from a few thousands to one hundred thousands sites. Each alignment consists of three sequences representing three primate species: human, chimp, and macaque. In our research, we focus on human and chimp sequences.

In order to better understand the real data, let us first look at some descriptive statistics. Figure 5.1 shows the histogram of the lengths of the sequence alignments. We note that the length of the most sequences falls within the range of 80000 to 300000.

We then examined the distribution of nucleotides in the sequences (Figure 5.2). Re-

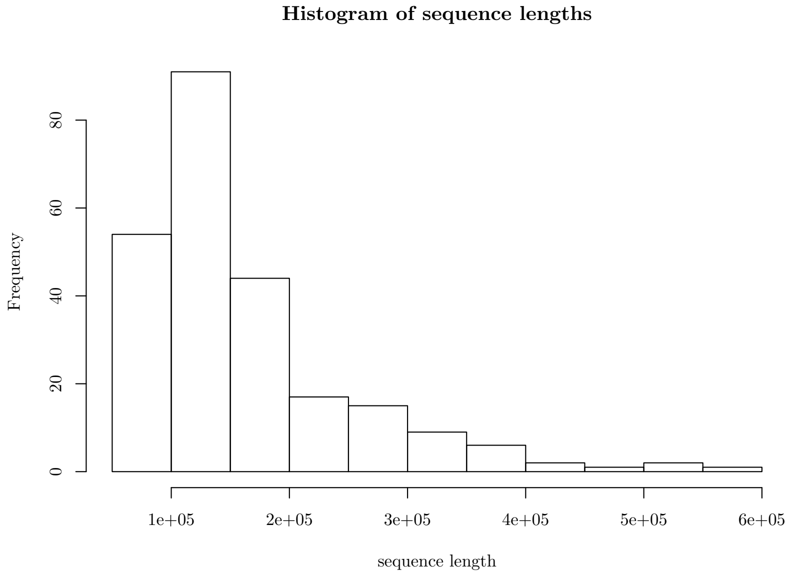

Figure 5.1: Histogram of sequence lengths

sults show that in most of the sequences, the percentages of the four types of nucleotides are between 15% to 35%. The median values of *T* and *A* types are relatively larger than those of *C* and *G* types.

We also determined the number of the substitutions that occurred in each pair of sequences. Figure 5.3 shows the histogram of the percentage of substitutions among all sites in the sequences. We note that the percentage of substitution lies between 0.2% to 1%.

Finally, we investigated the number of the substitutions that occurred in context dependent sites in each pair of sequences. Figure 5.4 shows the boxplots of the percentage of changes in context dependent sites. We see that when the left site is *C* or the right site is *G*, the percentage of changes is generally higher. This observation motivates us

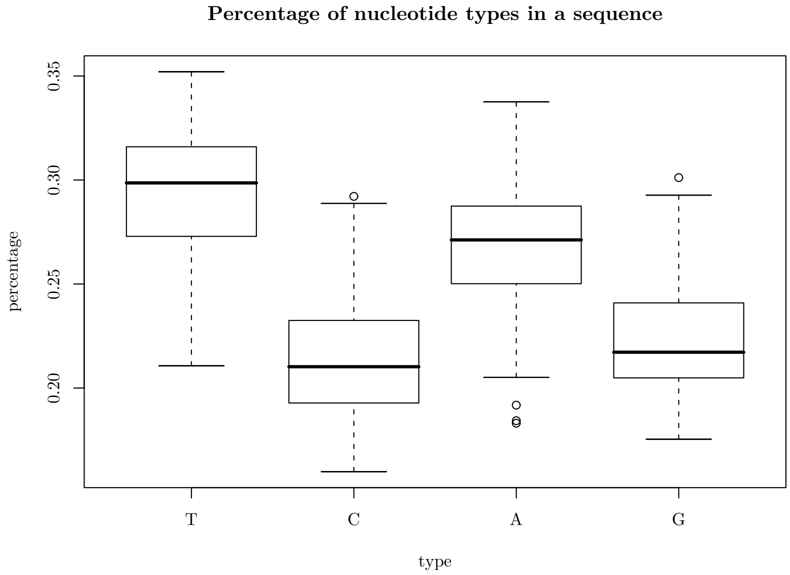

Figure 5.2: Percentage of nucleotide types in a sequence

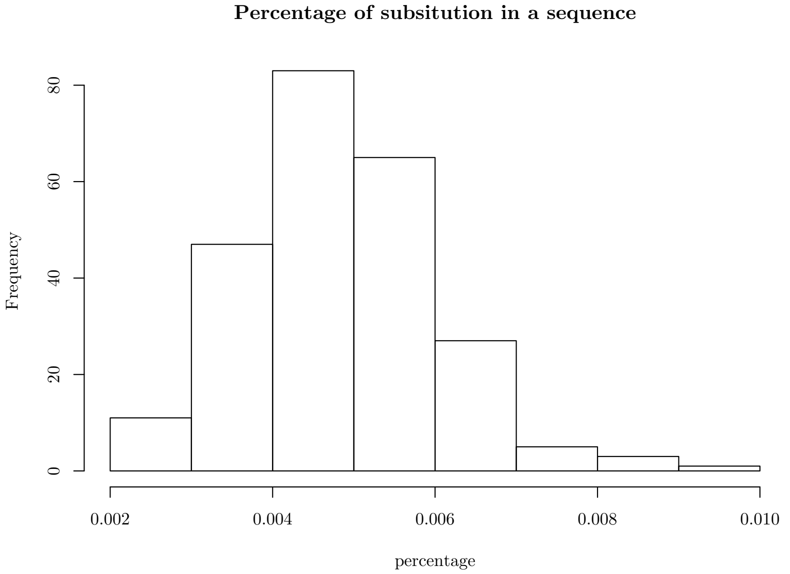

Figure 5.3: Histogram of percentage of substitution in a sequence

to consider a context dependent instead of a site independent model.

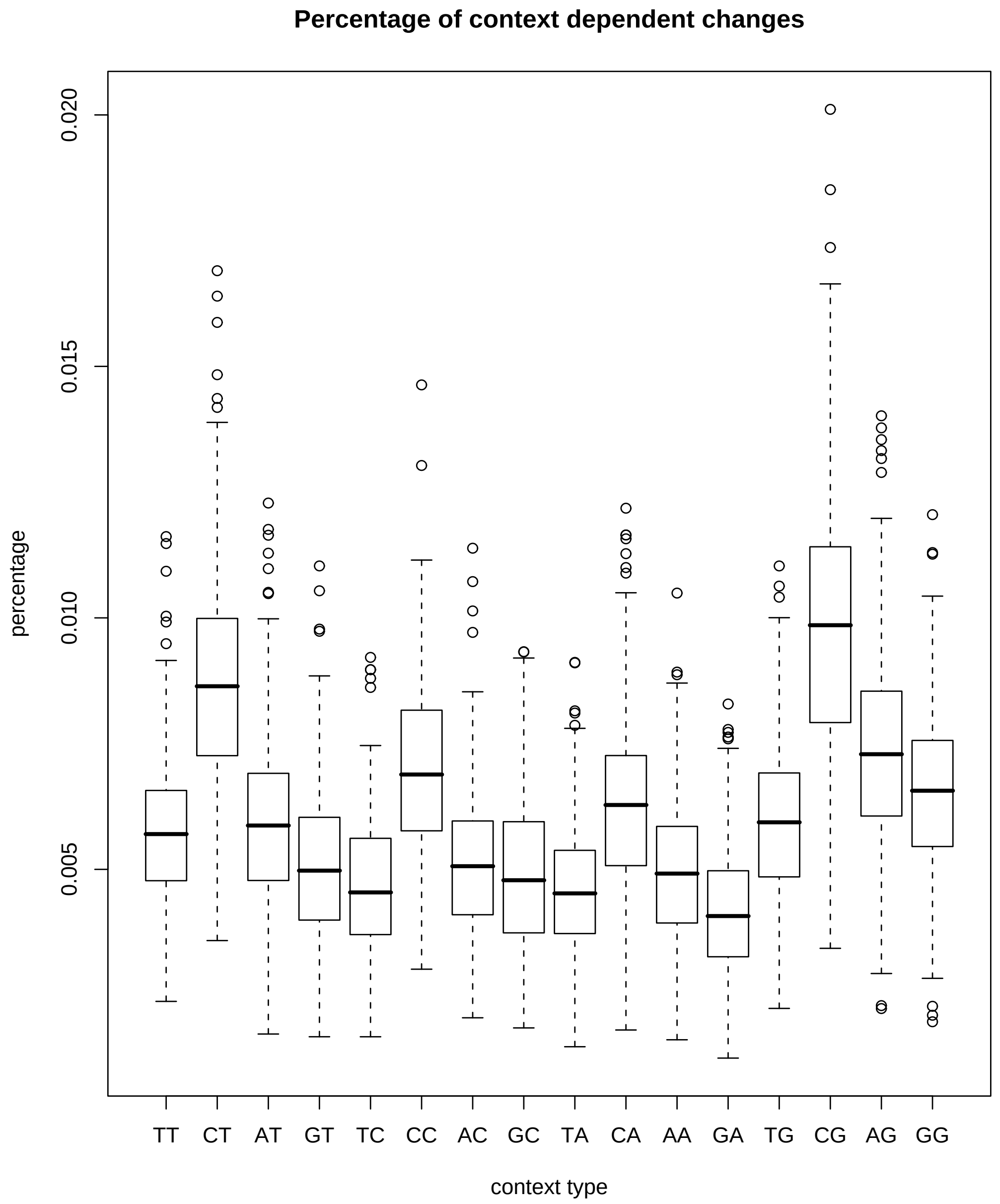

Figure 5.4: Percentage of changes in context dependent sites

5.2 Clustering of rate matrices

In the general two-flanking site context dependent model, 16 substitution rate matrices are defined for 16 context dependent cases. For each substitution matrix, there are 12 independent elements (the off diagonal elements). Therefore, each matrix is represented by a 12-dimensional vector, and 16 substitution rate matrices can be represented by a 12×16 matrix.

As the real data set consists of 242 samples, there are 242 12×16 matrices; Binding the matrices together, yields a large matrix of 12×242 rows and 16 columns. We used this data matrix for tree clustering. During clustering, the tree clustering approach joins two most similar clusters and form a new cluster in each step. This process is repeated until all the clusters become one single cluster. The final tree structure shows the relationship between the clusters.

Figure 5.5 shows the tree clustering plot of the matrices. Each initial cluster in this figure is labeled with a name, which means the context of the matrix, e.g. T_A means that the matrix is defined for sites whose left context is T and right context is A, and so on. From our clustering result, we can see that there are four clear clusters in the matrices. The first cluster is T_T, A_A, A_T, G_T, G_C, T_A, G_A, T_C, and A_C. The second cluster is C_T, C_C, and C_A. The third cluster is A_G, T_G, and G_G. The fourth cluster is C_G. The four clusters are exactly the same as our defined $4Q$ model earlier. This confirms the correctness of our definition of $4Q$ model. The $4Q$ matrices are Q_C , Q_G ,Q_{CG} and Q_{others}.

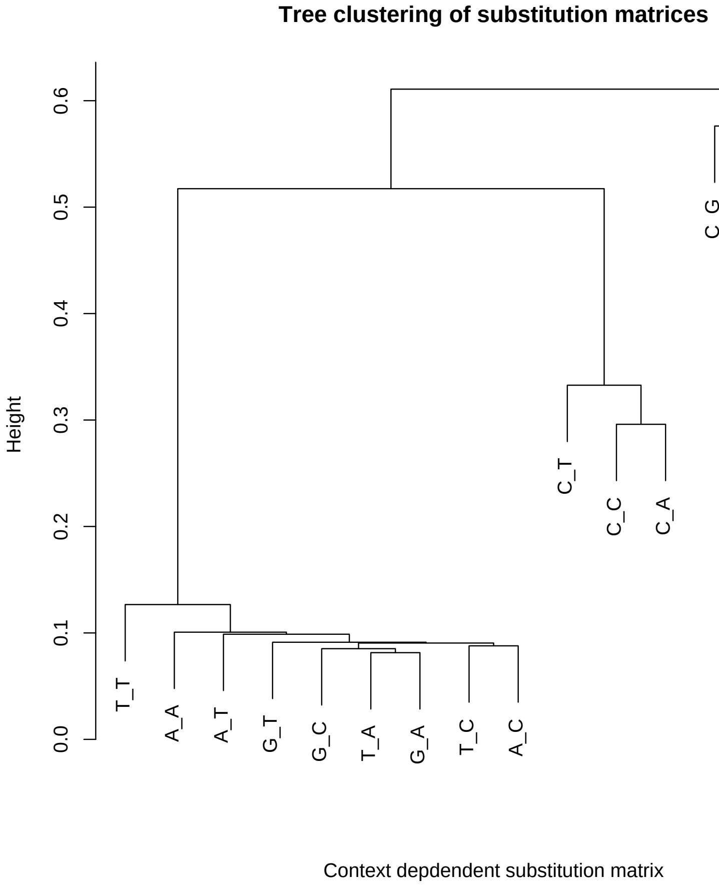

Figure 5.5: Tree clustering plot for rate matrices

5.3 Goodness of fit for the models

We performed goodness of fit to compare our models. In our study, we used the pseudo-likelihood to replace likelihood.

We made three comparisons to examine the difference between the general model and the sub-models.

5.3.1 Pseudo-likelihood values for different models

We first calculate the different pseudo-likelihood values for different models. Since we maximize the pseudo-likelihood to obtain the maximum estimates of the parameters, we expect that more parameters lead to larger pseudo-likelihood values.

Some sample results of Pseudo-likelihood values for different models for both branches are given in the Table 5.1. We see that $16Q$ model has the highest pseudo-likelihood values for both branches. The $4Q$ model come next; and $2Q$ model has the lowest pseudo-likelihood values.

Table 5.1: Pseudo-likelihood values for different models

	H branch			C branch		
Sample	16Q	4Q	2Q	16Q	4Q	2Q
1	-4064.398	-4198.563	-4347.697	-4003.212	-4106.676	-4292.987
2	-5952.966	-6375.546	-6447.132	-6476.525	-6834.293	-6929.349
3	-2164.779	-2279.275	-2325.443	-2231.921	-2348.025	-2406.109
4	-2901.398	-2980.173	-3068.342	-3143.730	-3319.023	-3387.515
5	-2802.812	-2940.759	-3026.644	-3038.313	-3168.272	-3262.058

5.3.2 $2Q$ model vs $4Q$ model

First, we performed the comparison of the $2Q$ model with the $4Q$ model. Our null hypothesis is $2Q$ model, and our alternative hypothesis is the $4Q$ model.

We used the pseudo-likelihood method to estimate the two models and to calculate the likelihood ratio scores of the two models. We then plotted histograms to show the likelihood ratio scores for the two branches (Figures 5.6 and 5.7).

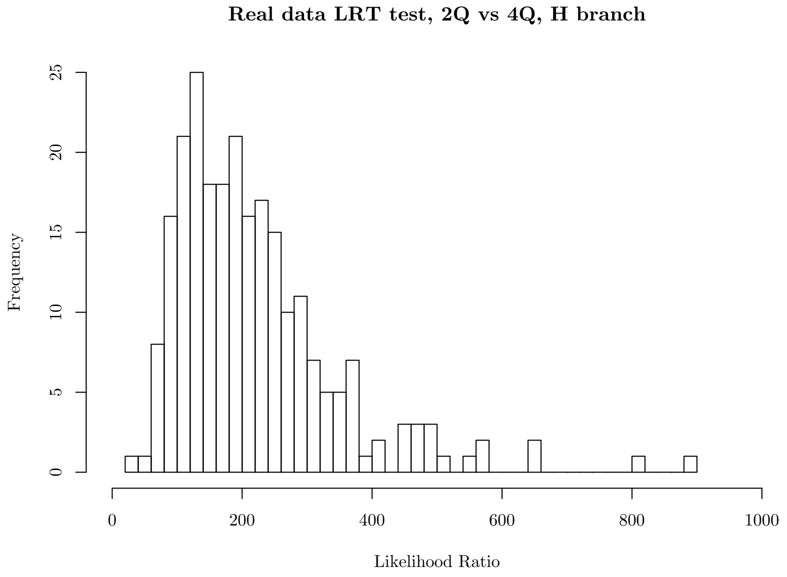

Figure 5.6: LRT test of 2Q vs 4Q for H branch

In chapter 4, from the simulation result, we know that the critical value of Test *A* for 5% significance level should be 40.5. From our calculation, we know that there are 99.6% LRT scores of *H* branch and 100% LRT scores of *C* branch are greater than 40.49. This indicates that the goodness-of-fit of 4Q is significantly better than 2Q for almost all alignments in our real data.

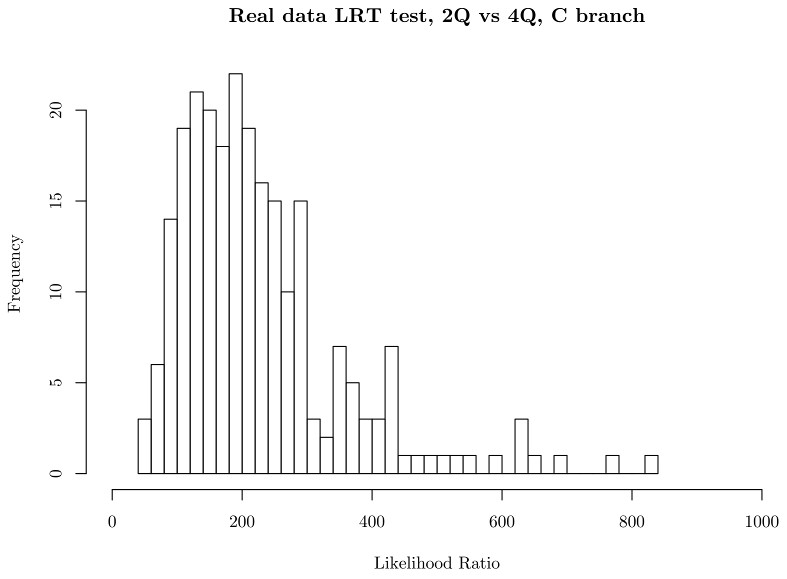

Figure 5.7: LRT test of 2Q vs 4Q for C branch

5.3.3 $2Q$ model vs $16Q$ model

Next, we compared the $2Q$ model with the $16Q$ general model. Our null hypothesis is the simple model, $2Q$ model, and our alternative hypothesis is the $16Q$ general model.

We used the pseudo-likelihood method to estimate the two models and to calculate the likelihood scores of the two models. We then plot histograms to show the likelihood ratio scores for the two branches (Figures 5.8 and 5.9).

Similarly, from the simulation result, we know that the critical value of Test *B* for 5% significance level should be 221.4. From our calculation, we see that there are 66.53% LRT scores of *H* branch and 66.12% LRT scores of *C* branch are greater than 221.4. This indicates that the goodness-of-fit of 16Q is significantly better than 2Q for 66% alignments in our real data.

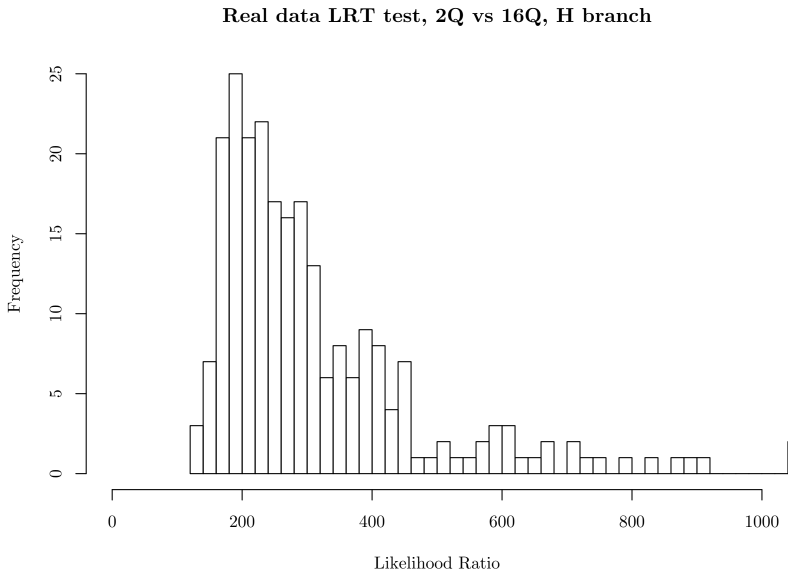

Figure 5.8: LRT test of 2Q vs 16Q for H branch

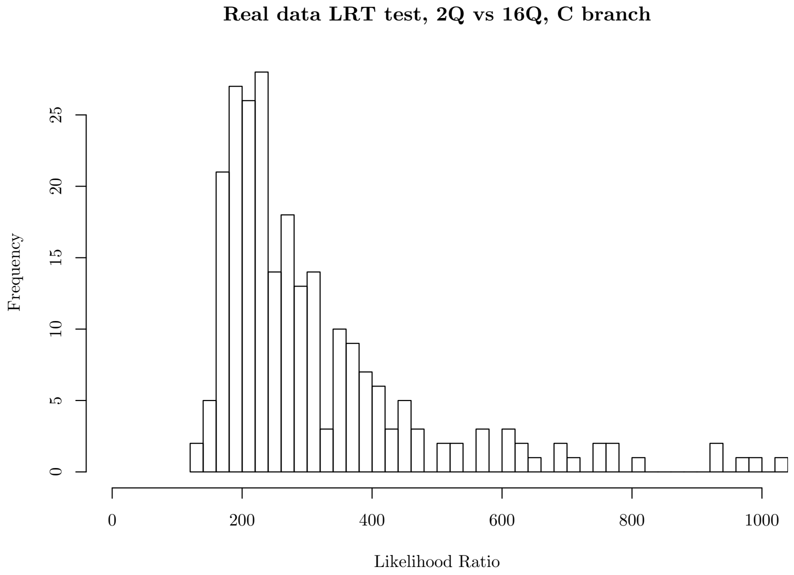

Figure 5.9: LRT test of 2Q vs 16Q for C branch

5.3.4 $4Q$ model vs $16Q$ model

Finally, we performed the comparison of the $4Q$ model with the $16Q$ general model. Our null hypothesis is the $4Q$ model, and our alternative hypothesis is the $16Q$ general model.

We used the pseudo-likelihood method to estimate the two models and to calculate the likelihood ratio scores of the two models. We then plotted histograms to show the likelihood ratio scores for the two branches (Figures 5.10 and 5.11).

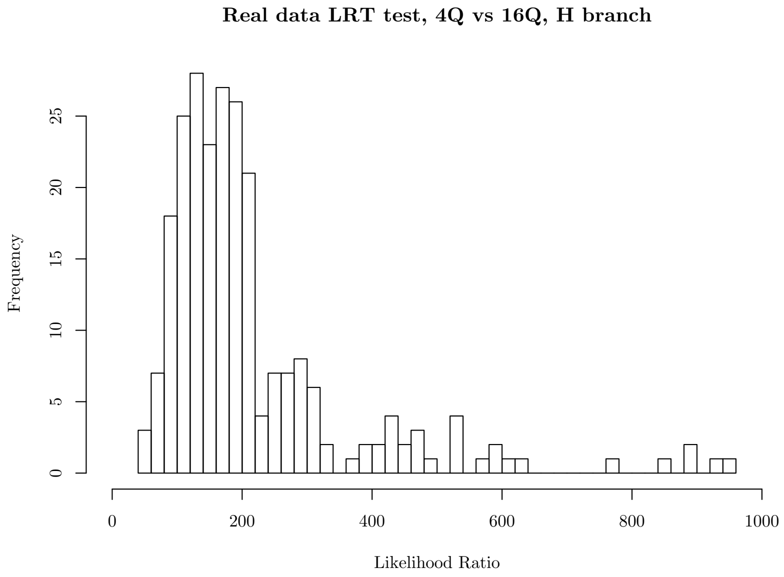

Figure 5.10: LRT test of 4Q vs 16Q for H branch

Similarly, from the simulation result, we know that maximum bound of the critical value of Test *C* for 5% significance level is 191.6. From our calculation, we know that there are 61.6% LRT scores of *H* branch and 66.1% LRT scores of *C* branch are less than 191.6. This indicates that the goodness-of-fit of 16Q is not significantly difference with 4Q for more than 60% alignments in our real data.

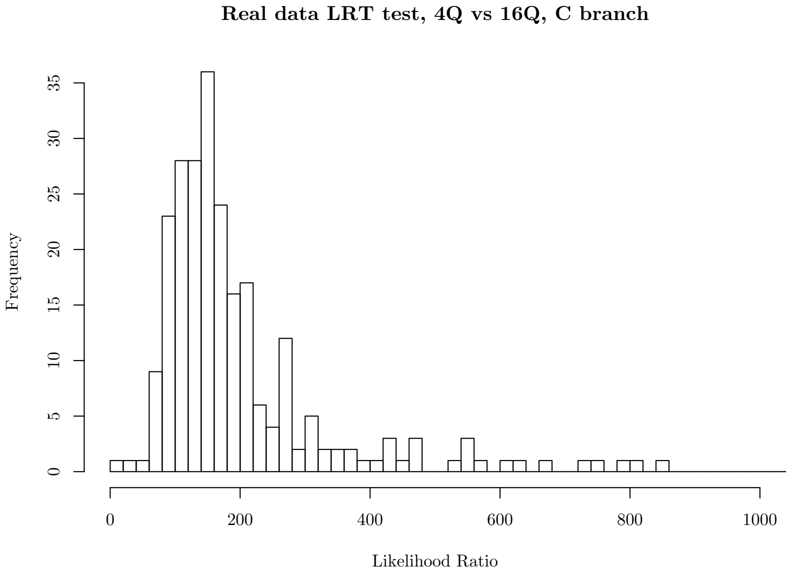

Figure 5.11: LRT test of 4Q vs 16Q for C branch

5.4 Summary

In this chapter, we applied the pseudo-likelihood method to a real data and conducted the goodness of fit test for our different models. Results show that the $2Q$ model is significantly different from $4Q$ model and the $16Q$ general model. But the $4Q$ model does not differ significantly from the $16Q$ general model. This indicates $4Q$ model is a good model to replace the $16Q$ general model for our application.

Chapter 6

Conclusion and further research

In this chapter, we summarize the work we have done and discuss some further research directions.

6.1 Conclusion

In the research of DNA sequence evolution, substitution rate matrices are used to describe the evolution process. When looking at the substitution of nucleotide, previous work normally ignore context dependence of the nucleotide. To better model the substitution process, context dependent substitution models need to be used. In this thesis, we have investigated the context dependent substitution rate models. Our work covered the following parts.

(1) Model definition

We proposed a general context dependent model framework, which used mathematical representation to describe the general cases of context dependent and independent models. Based on the general model, different context dependent models can be derived as special cases of the general model.

In the investigated special case, the two flanking sites context model, we used the neighboring sites of a nucleotide as the context. In the full model, 16 context dependent rate matrices are defined. Using clustering approach, we reduced the full model (16 matrices) into four matrices and two matrices as two simplified submodels.

(2) Model simplification

In context dependent substitution models, multiple substitution matrices were used for different context. This inevitably introduces many parameters. Previous works tried to reduce the number of parameters by reducing the number of independent parameters in each substitution matrix. However, we proposed to reduce the number of matrices based on knowledge of DNA evolution.

To reduce the number of matrices, the context cases need to be clustered into groups. In the thesis, we proposed to use statistical analysis method to cluster the context cases. This not only confirms the proposed submodels but also provided a general way for solving similar problems.

(3) Estimation methods

Previous works on context dependent models investigated the estimation of substitution rates from two known descendent sequences that are evolved from the same unknown ancestor sequence. Little research was done to estimate context dependent substitution rates from a given ancestor sequence and its descendent sequence. In our work, the rate estimation was based on the evolution from a known ancestor to a known descendent. We made use of the phylogenetic tree of the species to first estimate the the ancestor.

Parsimony approach is frequently used in the estimation of independent substitution models. In our work, we introduced it into the context dependent case. Also, we used a counting method to solve the problem of the changes of adjacent sites in DNA sequence. It overcomes the inaccuracy of standard methods when dealing with adjacent changes in DNA evolution.

We used optimization method for maximum pseudo-likelihood approach to estimate the substitution rates. The optimization process is very slow when the initial values are not properly given. Therefore, we proposed to use the rates estimated from the parsimony method as the initial values. This reduces the convergence time and increases the optimization speed.

(4) Simulation process

Previous research normally worked on limited real data. In our work, we developed a process to simulate context dependent DNA sequence evolutions. This provides us a flexibility of doing various experiment on simulated data.

The process simulated the context dependent substitution from given rate matrices and an initial sequence. We used simulation to evaluate different estimation methods.

(5) Evaluation methods

In the evaluation of different models, we proposed to use pseudo-likelihood ratio test to test the goodness of fit. We calculated the rate matrices for the real data using different models. We then compared the results by different models.

Major findings from our work are as follows.

(1) We used $2Q$ model and $4Q$ model as our simplified submodels. When using clustering method to group the similar matrices, the clustering tree shows a clear grouping of the matrices. This confirms that the $2Q$ and $4Q$ models are proper submodels.

(2) One of the problem for the context dependent model is that the context may change before the site in consideration changes. We modified the parsimony method to make it work in this situation. Our experiment show that the improved method that considers the change of context improves the estimation accuracy of substitution rates.

(3) The parsimony method works as well as the pseudo-likelihood approach when the substitution rates of evolution process are small (at the level of 0.001). When the rates are high, the parsimony method does not work well as the pseudo-likelihood method.

(4) When substitution rates are small, both parsimony and pseudo-likelihood methods work equally well under $2Q$ model. But under $4Q$ model, the pseudo-likelihood

method is superior to the parsimony method. The reason for the difference is that parsimony method overlooks the intermediate substitution process, and when substitution happens more frequently, it will get worse. This shows that the maximum pseudo likelihood methods is more robust.

(5) We applied the pseudo-likelihood method with different model definitions ($16Q$, $4Q$ and $2Q$ models) to the real data. From goodness-of-fit tests, $16Q$ is the most accurate model. The $2Q$ model has the smallest number of parameters. However, it has a fairly big difference in terms of likelihood ratio values compared to $4Q$ model and the $16Q$ general model. But the $4Q$ model does not differ much from the $16Q$ general model. This implies that the $4Q$ model is the best model for the real data as a comprise between the number of parameters and accuracy.

6.2 Further research topics

In our context dependent substitution model, we used a clustering approach to reduce the number of parameters of the model. That is, in our 16 matrices model, we used the clustering method to group the similar matrices and reduce the number of parameters. In independent substitution model, the number of independent parameters in substitution rate matrix, such as Jukes-Cantor, Kimura, HKY model and reversible model are reduced. By combining the matrices and simple models we may have more freedom to reduce the number of parameters while keeping the accuracy of the model.

In this thesis, we only considered the context dependent substitution model for nucleotide sequence. The methods may be extended to models for codon sequences.

References

Akaike, H. (1973). Information theory and an extension of the maximum likelihood principle. *2nd Int. Symp. on Information Theory,* 267-281.

Arndt, P. F., Burge, C. B. and Hwa, T. (2003a). DNA sequence evolution with neighbour-dependent mutation. *J. Comput. Biol.* **10**, 313-322.

Arndt, P. F., Petrov, D. and Hwa, T. (2003b). Distrinct changes of genomic biases in nucleotide substitution at the time of mammalian radiation. *Mol. Biol. Evol.* **20**, 1887-1896.

Arndt, P. F. and Hwa, T. (2005). Identification and measurement of neighbor-dependent nucleotide substitution processes. *Bioinformatics.* **21**, 2322-2328.

Becker, R. A., Chambers, J. M. and Wilks, A. R. (1988) The new S language. *Wadsworth and Brooks Cole.*

Besag, J. (1975). Statistical analysis of non-lattice data. *The Statistician.* **24**, 179-195.

Broyden, C. G.(1970). The Convergence of a Class of Double-rank Minimization Algorithms. *Journal of the Institute of Mathematics and Its Applications* **6**, 76-90

Camin, J.H. and Sokal, R. R. (1965). A method for deducting branching sequences in phylogeny. *Evolution.* **19**, 311-326.

Christensen, O.F., Hobolth, A. and Jensen, J.L. (2005). Pseudo-likelihood analysis of codon substitution models with neighbor dependent rates. *J. Comput. Biol.* **12**, 1166-1182.

Christensen, O.F. (2006). Pseudo-likelihood for non-reversible nucleotide substitution models with neighbour dependent rates. *Statistical Applications in Genetics and Molecular Biology.* **5**, iss1, art18.

Deonier, R. C., Tavare, S. and Waterman, M. S. (2005). Computational Genome Analysis An Introduction. *Springer.*

Durbin, R., Eddy, S., Krogh, A., and Mitchison, G. (1998). Biological sequence analysis: Probabilitstic models of proteins and nucleic acids. *Cambridge University Press, London.*

Farris, J.S. (1970). Methods for computing Wagner trees. *Systematic Zoology.* **19**, 83-92.

Felsenstein, J. (1973). Maximum-likelihood estimation of evolutionary trees from continuous characters. *J. Am J Hum Genet.* **25**, 471-492.

Felsenstein, J. (1978). Cases in which parsimony or compatibility methods will be positively misleading. *Syst. Zool.* **27**, 401-410.

Felsenstein, J. (1981a). Evolutionary trees from DNA sequences: a maximum likelihood approach. *J. Mol. Evol.* **17**, 368-376.

Felsenstein, J. (1981b). A likelihood approach to character weighting and what it tells us about parsimony and compatibility. *Biol. J. Linn. Soc.* **16**, 183-196.

Felsenstein, J. (1988). Phylogenies from molecular sequences: inference and reliability. *J. Annu Rev Genet.* **22**, 521-565.

Felsenstein, J. (1996). Inferring phylogenies from protein sequences by parsimony, distance, and likelihood methods. *J. Methods Enzymol.* **266**, 418-27.

Felsenstein, J. and Churchill, G. A. (1996). A Hidden Markov Model approach to variation among sites in rate of evolution. *Mol Biol Evol.* **13**, 93-104.

Felsenstein, J. (2004). Inferring Phylogenies. *Sinauer Associates, Inc., Sunderland, Massachusetts.*

Fletcher, R.(1970). A New Approach to Variable Metric Algorithms. *Computer Journal* **13**, 317-322

Fowlkes, E. B. and Mallows, C. L. (1983). A Method for Comparing Two Hierarchical Clusterings. *Journal of the American Statistical Association* **78**, 553C584.

Geys, H., Molenberghs, G. and Ryan, L. (1999). Pseudo-likelihood modelling of multivariate outcomes in developmental toxicology. *Journal of the American Statistical Association* **94**, 734-745.

Graur, D. and Li, W.H. (2000). Fundamentals of Molecular Evolution: Second Edition*Sinauner Associates INC. Publishers, Sunderland, Massachusetts.*

Goldman, N. and Yang, Z. (1994). Models of DNA substitution and the discrimination of evolutionary parameters. *In Proceedings of the XVIIth International Biometrics Conference.* **I**, 407-420.

Goldman, N. and Yang, Z. (1994). A condon-based model of nucleotide substitution for protein-coding DNA sequences. *Mol. Biol. Evol.* **11**, 725-736.

Goldfarb, D.(1970). A Family of Variable Metric Updates Derived by Variational Means. *Mathematics of Computation* **24**, 23-26

Gojobori T, Li W.H., Graur D. (1982b). Patterns of nucleotide substitution in pseudogenes and functional genes. *J. Mol. Evol.* **18**, 360-369.

Hasegawa, M., kishino, H. and Yano, T.(1985). Dating of the human-ape splitting by a molecular clock of mitochondrial DNA. *J. Mol. Evol.* **22**, 160-174.

Hobolth, A. (2008). A Markov Chain Monte Carlo Expectation Maximization algorithm for statistical analysis of DNA sequence evolution with neighbor-dependent substitution rates. *Journal of Computational and Graphical Statistics.* **17**, 138-162.

Huelsenbeck, J. P. and Crandall, K. A.(1997). Phylogeny estimation and hypothesis testing using maximum likelihood. Ann. *Rev. Ecol. Syst.* **28**, 437-466.

Huelsenbeck, J. P. and Rannala, B.(1997). Phylogenetic methods come of age: testing hypotheses in an evolutionary context. *Science* **276**, 227-232.

Huttley, G.A. (2004). Modeling the impact of DNA methylation on the evolution of BRCA1 in mammals. *Mol. Biol. Evol.* **21**, 1760-1768.

Hwang, D. and Green, P.(2004). Bayesian Markov chain Monte Carlo sequence analysis reveals varying neutral substitution patterns in mammalian evolution. *PNAS* **101**, 13994-14001.

Jensen, J. L. (2005). Context dependent DNA evolutionary models. *Research Report* **458**, Department of Theoretical Statistics, Aarhus University.

Jensen, J. L. and Pedersen, A. -M. K.(2000). Probabilistic models of DNA sequence evolution with context dependent rates of substitution. *Adv. Appl. Prob* **32**, 499-517.

Johnson, N. L., Kotz, S. and Balakrishnan, N. (1995). Continuous Univariate Distributions chapters 18 (volume 1) and 29 (volume 2). *Wiley, New York.*

Johnson, R. A. and Wichern, D. W. (2002). Applied multivariate statistical analysis. (Fifth Edition) *Prentice Hall.*

Juckes, T. and Cantor, C. (1969). Evolution of protein molecules. In H. Munro (Ed) *Mammalian Protein Metabolism* **3**, 21-132. Academic Press, New York.

Karlin, S. and Burge, C. (1995). Dinucleotide relative abundance extremes: a genomic signature. *Trends Genet.* **11**, 283-290.

Kelly, F. P. (1979). Reversibility and stochastic networks. *John Wiley and Sons, New York.*

Kimura, M.(1980). A simple method fro estimating evolutionary rates of base substitutions through comparative studies of nucleotide sequences. *J. Mol. Evol.* **16**, 111-120.

Kishino, H. and Hasegawa, M. (1989). Evaluation of the maximum likelihood estimate of the evolutionary tree topologies from DNA sequence data, and the branching order in Hominoidea. *J. Mol. Evol.* **29**, 170-179.

Krawczak M., Ball E.V., Cooper D.N. (1998). Neighboring nucleotide effects on the rates of inherited single base-pair substitution in human genes. *American Journal of Human Genetics.* **63**, 474-488.

Lunter, G. and Hein, J. (2004). A nucleotide substitution model with nearest-neighbour interations. *Bioinformatics* **20**, i216-i223.

Neyman, J. and Pearson, E.S (1967). On the Use and Interpretation of Certain Test Criteria for Purposes of Statistical Inference, Part I. *Cambridge University Press, Cambridge*.pp.1-66

Neyman, J. and Pearson, E.S.(1967). The testing of statistical hypotheses in relation to probabilities a priori. *ambridge University Press, Cambridge*. pp.186-202

Pearson, E.S. and Neyman, J. (1967). On the Problem of Two Samples. *ambridge University Press, Cambridge*. pp.99-115.

Pedersen, A.-M. K. and Jensen, J. L.(2001). A dependent-rates model and an MCMC-based methodology for the maximum-likelihood analysis of sequences with overlapping reading frames. *Mol. Biol. Evol.* **18**, 763-776.

Shimodaira, H. and Hasegawa, M.(1999). Multiple comparisons of log-likelihoods with applications to phylogenetic inference. *Mol. Biol. Evol.* **16**, 1114-1116.

Shanno, D. F.(1970). Conditioning of Quasi-Newton Methods for Function Minimization. *Mathematics of Computation* **24**, 647-656.

Siepel, A. and Haussler, D.(2004). Phylogenetic estimation of context-dependent substitution rates by maximum likelihood. *Mol. Biol. Evol.* **21**, 468-488.

Swofford, D. L., Olsen, G. J., Waddell, P. J. and Hillis, D. M. (1996). Phylogenetic inference. Pp. 407-543 in Hillis, D. M., Moritz, C. and Mable, B. K. eds., Molecular Systematics, second edition. *Sinauer Associates, Sunderland, Massachusetts.*

Tavare, S.(1986). Some Probabilistic and Statistical Problems in the Analysis of DNA Sequences. *Mathematics in the Life Science* **17**, 57-86.

Whelan, S. and Goldman, N. (2004). Estimating the Frequency of Events That Cause Multiple-Nucleotide Changes. *Genetics.* **167**, 2027-2043.

Yap, V. B. and Speed, T. P. (2004). Modeling DNA base substitution in large genomic regions from two organisms. *J. Mol. Evol.* **58**, 12-18.

Yap, V. B. and Speed, T. P. (2005). Estimating substitution matrices. In: *Statistical Methods in Molecular Evolution* (ed. R. Nielsen), Springer, New York, 407-438.

Yang, Z.(1993). Maximum likelihood estimation of phylogeny from DNA sequences when substitution rates differ over sites. *J. Mol. Evol.* **10**, 1396-1401.

Yang, Z.(1994). Statistical properties of the maximum likelihood method of phylogenetic estimation and comparison with distance matrix method. *Syst. Biol.* **43**, 329-342.

Yang, Z.(1994). Maximum likelihood phylogenetic estimation from DNA sequences with variable rates over sites. *J. Mol. Evol.* **39**, 306-314.

Yang, Z., Goldman, N. and Friday, A. E. (1994). Comparison of models for nucleotide substitution used in maximum likelihood phylogenetic estimation. *Mol. Biol. Evol.* **11**, 316-324.

Yang, Z.(1994). Estimating the pattern of nucleotide substitution. *J. Mol. Evol.* **39**, 105-111.

Yang, Z. and Goldman, N. (1994). Evaluation and extension of Markov process models for the evolution of DNA. *Acta Genetica Sinica* **21**, 17-23.

Yang, Z.(1994). Estimating the pattern of nucleotide substitution. *J. Mol. Evol.* **39**, 105-111.

Yang, Z.(1995). Evaluation of several methods for estimating phylogenetic trees when substitution rates differ over nucleotide sites. *J. Mol. Evol.* **40**, 689-697.

Yang, Z.(1995). A space-time process model for the evolution of DNA sequences. *Genetics* **139**, 993-1005.

Yang, Z., Goldman, N. and Friday, A. E.. (1995). Maximum likelihood trees from DNA sequences: a peculiar statistical estimation problem. *Syst. Biol.* **44**, 384-399.

Yang, Z.(1996). Phylogenetic analysis using parsimony and likelihood methods. *J. Mol. Evol.* **42**, 294-307.

Yang, Z.(1996). Maximum likelihood models for combined analyses of multiple sequence data. *J. Mol. Evol.* **42**, 587-596.

Yang, Z.(1997). PAML: a program for package for phylogenetic analysis by maximum likelihood. *CABIOS* **15**, 555-556.

Yang, Z.(1998). Likelihood ratio tests for detecting positive selection and application to primate lysozyme evolution. *Mol. Biol. Evol.* **15**, 568-573.