Inference of Rates across Sites via an Expectation Maximization Algorithm

by

Tingting Zhao

B.Sc., Zhejiang University, 2011

A THESIS SUBMITTED IN PARTIAL FULFILLMENT OF
THE REQUIREMENTS FOR THE DEGREE OF

MASTER OF SCIENCE

in

The Faculty of Graduate Studies

(Statistics)

THE UNIVERSITY OF BRITISH COLUMBIA

(Vancouver)

September 2013

Abstract

The rates of nucleotide substitution can be different from genes to genes. Moreover, different regions of the same gene can have different rates of mutation as well. Many attempts have been tried to allow for the variable rates across different nucleotide sites. A rate factor coming from the continuous distribution has been introduced to deal with the problem. However, for computation reasons, this method can only scale to less than a dozen sequences. Later studies use a discrete gamma distribution to approximate the gamma distribution.

The main contribution of our work is that we propose a discrete distribution over the rate factor which is more flexible while preserving attractive computational properties. We make inference about the rate factor and its distribution via an Expectation Maximization (EM) algorithm. We evaluate our method by both simulations and a real dataset. From the real dataset, it reflects that the method is useful for large phylogenies with even thousands of sequences. We analyze the identifiability of our model for a pair of DNA sequences under certain conditions. We also prove for certain types of rate matrices, this model is non-identifiable.

Preface

This dissertation is original, unpublished, independent work by the author, Tingting Zhao.

Table of Contents

Appendices

List of Tables

List of Figures

Acknowledgements

First and foremost I would like to express my deepest gratitude to my supervisor Dr. Alexandre Bouchard-Côté for his guidance, insights and encouragement for the last two years. I feel lucky to work with him since he has taken me to start a wonderful journey in the research world. I am influenced by his enthusiasm and attitude for research and work. I would like to thank him for his dedication in revising the thesis. I am also grateful to Dr. Lang Wu for being my second reader and his precious suggestions and comments.

Second, I would like to thank my peers including Yanling(Tara) Cai, Yongliang(Vincent) Zhai, Yi Huang, Yunlong Nie, Seong-Hwan Jun, Tingting Yu and Andy Leung for sharing their research experiences and discussions with me.

Last, I would like to thank my parents for bringing me to this world and always trying their best to help me realize my dreams. This thesis will serve as a gift to them.

Chapter 1

Introduction

1.1 Motivation

Phylogenetics is the study of the evolutionary relationships among groups of organisms. Continuous time Markov chains (CTMC) are at the core of modern phylogenetic methods to model the evolutionary process of DNA sequences. Several different DNA models have been proposed such as the JC69 Model by Jukes and Cantor (1969), the K80 Model by Kimura (1980) and so on. If DNA sequences are available for several species, we can make inference about the phylogeny of these species via the maximum likelihood method and obtain estimates of the parameters such as the topology of the tree, the transition matrix and the branch lengths.

If there is no rate variation, all sites share the same rate matrix Q describing the instantaneous rate of different kinds of substitutions with different bases. However, it has been discovered that the mutation rates of different regions of the same gene can be different as shown by Graur and Li (2000). To account for the rate variation over sites, several approaches have been proposed. Yang (1993) proposed a rate factor coming from the gamma distribution. A rate factor γ is a parameter assigned to each site to adjust the rate variation by multiplying it to the rate matrix Q for this site. However, for computational reasons, this method can only scale to less than a dozen sequences. To simplify the computation, Yang (1994) uses the "discrete gamma distribution" to approximate the continuous gamma distribution, but he points out there is overestimation or underestimation of the shape parameter of the gamma distribution given different tree topologies. Moreover, there is no direct biological reasons to favour the gamma distributions of the rate factor. We propose a new model to allow the rate variation which is more flexible while preserving attractive computational properties.

We assume a discrete distribution over the rate factor without other restrictions. We make inference about the rates and their distribution via an EM algorithm for both pairs of sequences as well as trees with large phylogenies. We evaluate the method with a real dataset. It reflects that the method is practical for large phylogenies with even thousands of sequences.

When doing simulations to check whether the EM algorithm can recover the rate factor and its distribution used to generate the dataset, we find that different rate factors and distributions can give a similar likelihood of the same generated dataset. We are motivated to study the identifiability of the model. In the context of phylogenetics, a model is non-identifiable if different set of parameters including tree topologies, branch lengths and evolutionary parameters can produce the same likelihood. Many people contributed to investigate the identifiability of models with the rate factor coming from a continuous distribution like the gamma distribution with mean one. Steel (2009) used the F81 Model and discovered that the shape parameter of the gamma distribution and the topology of the tree are not identifiable. Wu and Susko (2010) proved the identifiability of general time reversible (GTR) models. Allman, Ané, and Rhodes (2008) proved that the four-state GTR + Γ model is identifiable given the joint distribution of at least triples of taxa.

However, few attempts are done in the literature to study the identifiability of the rate scalar from a discrete distribution.

The reason leads to that is the difficulty in obtaining the solution of the rate factor and its distribution of the inverse moment generating function. For the gamma distribution with mean one, its inverse moment generating function is only determined by the shape parameter. In our work, we prove the non-identifiability of the JC69 Model and the F81 Model of a pair of DNA sequences in the context of four distinct category modes. We also prove the non-identifiability of other DNA evolution models with two or three distinct eigenvalues under certain conditions. The identifiability of the models under tree structures is still an open question.

1.2 Outline

In Chapter 2, we review some popular Markov models of DNA sequence evolution in the framework of CTMC. After introducing the rate matrices for different DNA evolution models, we illustrate how to deal with unequal evolution rates of different sites.

In Chapter 3, we introduce how to calculate the likelihood for a pair of homologous DNA sequences without rate variations and also with rate heterogeneity. Moreover, we also provide how to calculate the likelihood of a given tree in those two situations.

In Chapter 4, we propose a discrete distribution of the rate factor to deal with rate heterogeneity. We explain how to make inference of the rate factor and its distribution via an EM algorithm for a pair of homologous DNA sequences and generalize it to a tree.

In Chapter 5, we summarize the previous work and results about the identifiability of models assuming the rate factor coming from a gamma distribution. We also analyze the identifiability of our model assuming a discrete distribution. We prove the non-identifiability of the JC69 Model and the F81 Model of a pair of DNA sequences in the context of four distinct category modes. We also prove the non-identifiability of other DNA evolution models under certain conditions. The identifiability of the models under tree structures is still an open question for future work.

In Chapter 6, we evaluate our EM algorithm with a real dataset with 1028 DNA sequences. It reflects that our algorithm is computationally attractive for trees with large phylogenies.

In the conclusion and future work part, we summarize the results of the thesis and discuss possible future work.

Chapter 2

Introduction to common DNA Evolution Models

2.1 Introduction

In this chapter, we review some popular Markov models of DNA sequence evolution in the framework of continuous time Markov chains (CTMC). Since the time of divergence between different pairs of homologous DNA sequences descending from a common ancestral sequence can widely vary, different branch lengths are introduced to represent the expected number of nucleotide substitutions between sequences. Time homogeneity is assumed in the continuous time Markov chains model so that the instantaneous rate matrix can be used to describe the substitution process. The difference between these models lies in the parameters describing the rates of different substitutions. The JC69 Model assumes equal transition rates, for example, the probability of a nucleotide A to change into other nucleotides G, T or C is the same while the K80 Model considers that the probability between purines such as a nucleotide A to change into G is larger than the changes between purines and pyrimidines such as a nucleotide A to change into T or C because of the similarity in the structure of A and G.

After introducing the rate matrices for different models, we also introduce how to deal with rates variations of different sites by introducing a rate factor multiplying the rate matrix of each site proposed by Yang (1993). To simplify the calculation, Yang (1994) uses the “discrete gamma distribution” model to approximate the continuous gamma distribution.

2.2 Models and Data Structure

2.2.1 Data Structure

The data comes from DNA sequences from homologous regions for species $i \in \{1, 2, \ldots, n_1\}$. Let $X = \{x_{ij}\}$ be the aligned nucleotide sequences, where $i \in \{1, 2, \ldots, n_1\}$, $j \in \{1, 2, ..., n_2\}$. Then n_2 is the number of nucleotides per sequence. Each column of the data matrix $x_j = \{x_{1,j}, ..., x_{n_1,j}\}$ specifies the nucleotides for the n_1 sequences at the jth site. The site is the position of a nucleotide in DNA sequences. Each row of the data matrix $x_i = \{x_{i,1}, ..., x_{i,n_1}\}$ represents all the nucleotides of the ith DNA sequence.

$$X = \begin{pmatrix} x_{1,1} & x_{1,2} & \cdots & x_{1,n_2} \\ x_{2,1} & x_{2,2} & \cdots & x_{2,n_2} \\ \vdots & \vdots & \ddots & \vdots \\ x_{n_1,1} & x_{n_1,2} & \cdots & x_{n_1,n_2} \end{pmatrix}$$

2.2.2 Model of DNA Evolution in the framework of CTMC

In the framework of CTMC, $\Omega = \{A, G, T, C\}$ represents the state space consisting four kinds of nucleotides in DNA sequences. Each individual entry refers to the probability that the state i will change into the state j, where $i, j \in \Omega$. $P(t)$ is the transition matrix, where t is the branch length representing the expected number of nucleotide substitution for a pair of homologous DNA sequences descending from a common ancestral sequence.

$$P(t) = \begin{pmatrix} p_{AA}(t) & p_{AG}(t) & p_{AC}(t) & p_{AT}(t) \\ p_{GA}(t) & p_{GG}(t) & p_{GC}(t) & p_{GT}(t) \\ p_{CA}(t) & p_{CG}(t) & p_{CC}(t) & p_{CT}(t) \\ p_{TA}(t) & p_{TG}(t) & p_{TC}(t) & p_{TT}(t) \end{pmatrix}$$

The instantaneous rates of change from one state to another is reflected by Q where $Q = \frac{dP(t)}{dt}$ with $P_0 = I$. In turn, $P_t = e^{tQ} = \sum_j \frac{(tQ)^j}{j!}$,

$$Q = \begin{pmatrix} * & \theta_{AG} & \theta_{AT} & \theta_{AC} \\ \theta_{GA} & * & \theta_{GT} & \theta_{GC} \\ \theta_{TA} & \theta_{TG} & * & \theta_{TC} \\ \theta_{CA} & \theta_{CG} & \theta_{CT} & * \end{pmatrix}.$$

The diagonal elements are specified to make sure the sum of each row in the Q matrix is zero. The Q matrix will also be constrained by multiplying each element of the matrix by a same factor $\mu = -1/\sum_{i\{A,C,G,T\}} \pi_i Q_{ii}$, where π_i is the stationary distribution of the rate matrix. This normalization ensures a branch length of one yields one expected change per nucleotide.

2.2.3 Rate Matrices of Different Models

JC69 Model proposed by Jukes and Cantor (1969)

$$Q = \begin{pmatrix} * & \frac{\mu}{4} & \frac{\mu}{4} & \frac{\mu}{4} \\ \frac{\mu}{4} & * & \frac{\mu}{4} & \frac{\mu}{4} \\ \frac{\mu}{4} & \frac{\mu}{4} & * & \frac{\mu}{4} \\ \frac{\mu}{4} & \frac{\mu}{4} & \frac{\mu}{4} & * \end{pmatrix}$$

In this model, the stationary distribution is $\pi = (\frac{1}{4}, \frac{1}{4}, \frac{1}{4}, \frac{1}{4})$ and μ is the standardization factor, where $\mu = -\frac{4}{3}$.

K80 Model proposed by Kimura (1980)

$$Q = \begin{pmatrix} * & \kappa & 1 & 1 \\ \kappa & * & 1 & 1 \\ 1 & 1 & * & \kappa \\ 1 & 1 & \kappa & * \end{pmatrix}$$

In this model, the standardization factor $\mu = 1/4\,(\kappa + 2)$. The κ in the matrix is the ratio of transition and transversion which are two types of DNA substitution mutations. Transitions are interchanges between both purines including A and G or both pyrimidines including C and T. Transversions are interchanges between purines and pyrimidines. Purines are two-ring structure while pyrimidines are one-ring structures. As a result, it is typically assumed that transitions are more likely to happen than transversions i.e. $\kappa \geqslant 1$.

F81 Model proposed by Felsenstein (1981)

$$Q = \begin{pmatrix} * & \pi_C & \pi_A & \pi_G \\ \pi_T & * & \pi_A & \pi_G \\ \pi_T & \pi_C & * & \pi_G \\ \pi_T & \pi_C & \pi_A & * \end{pmatrix}$$

This model allows for different base frequencies for four different states A, G, T and C. The stationary distribution is $\{\pi_A, \pi_G, \pi_T, \pi_C\}$, where the standardization factor μ is $1/\big(1 - \pi_A^2 - \pi_C^2 - \pi_G^2 - \pi_T^2\big)$.

HKY85 Model proposed by Hasegawa, Kishino, and Yano (1985)

$$Q = \begin{pmatrix} * & \kappa\pi_C & \pi_A & \pi_G \\ \kappa\pi_T & * & \pi_A & \pi_G \\ \pi_T & \pi_C & * & \kappa\pi_G \\ \pi_T & \pi_C & \kappa\pi_A & * \end{pmatrix}$$

The model does not assume equal base frequencies for the four different states and accounts for the difference between transitions and transversions with one parameter κ in the rate matrix Q, where the normalization constant μ is $1/(2(\pi_A + \pi_G)(\pi_C + \pi_T) + 2\kappa(\pi_A\pi_G + \pi_C\pi_T))$.

In this thesis, we are using the HKY85 Model in the simulation study since it can incorporate rate variations considering different base frequencies and the bias in transitions over transversions shown by many genes. But in the real dataset, we are using the K80 model since it is more simple than HKY85 model and we assume a uniform distribution of the four states of $\{A, G, T, C\}$.

2.2.4 Models considering Rate Variation between Different Nucleotide Sites

According to Graur and Li (2000), the rate of nucleotide substitution τ is defined as the number of substitution per site per year. If no rate variation across nucleotide sites is introduced, the rates of substitution are the same for all sites on the DNA sequences according to the Q matrix. For example, assuming there are 1000 sites on a pair of DNA sequences, then on all these sites, the rate for a nucleotide to change from A to G is the same. However, this assumption may not hold. We cite a table from Graur and Li (2000) as

Table 2.1 to illustrate that the numbers of nucleotide substitutions per site (K) on regions of genes can be different.

Table 2.1: Numbers of nucleotide substitutions per site (K) between cow and goat $\beta-$ and $\gamma-$globin genes and between cow and goat $\beta-$globin pseudogenes cited from Graur and Li (2000)

Region	K^a
5' flanking region	5.3 ± 1.2
5' untranslated region	4.0 ± 2.0
Fourfold degenerate sites	8.6 ± 2.5
Introns	8.1 ± 0.7
3' untranslated region	8.8 ± 0.2
Pseudogenes	9.1 ± 0.9

The rates in the table are in units of substitutions per site per 10^9 years. From the table, it reflects that the rates of nucleotide substitution are different in different regions. These regions are classified according to different functions that they perform during transcription and translation. Transcription is the synthesis of a RNA molecule based on a DNA template and translation is the process of the produced RNA during transcription conveying information to the ribosomes to create proteins.

If the rates of change are different for distinct sites, then some sites evolve more quickly than others. In this situation, Yang (1993) proposed a rate factor γ_i for the ith site where γ_i comes from the gamma distribution.

Then the rate matrix for the ith site is

$$Q_i = \gamma_i \times \begin{pmatrix} * & \kappa\pi_C & \pi_A & \pi_G \\ \kappa\pi_T & * & \pi_A & \pi_G \\ \pi_T & \pi_C & * & \kappa\pi_G \\ \pi_T & \pi_C & \kappa\pi_A & * \end{pmatrix}$$

However, instead of making inference of γ_i for each site,Yang (1993) proposed that all possible rates could be integrated out for each site when calculating the likelihood which will be covered in the next chapter.

It has been noted that using the gamma distribution is very computationally expensive and in order to improve that, Yang (1994) proposed the "discrete gamma distribution" to approximate the continuous gamma distribution.

This project relaxes the assumption of equal rates of substitution for all sites by multiplying Q with one category of a rate factor $\gamma = (\gamma_1, \ldots, \gamma_k)$. However, we only assume γ comes from a discrete distribution.

Definition 1. *We define the rate factor $\gamma = (\gamma_1, \ldots, \gamma_k)$ with a discrete distribution $f = (f_1, f_2, \ldots, f_k)$ as a parameter assigned to each site to adjust the rate variation by multiplying one category of γ to the rate matrix Q for this site, where f_k denotes the probability for this site to take the kth category of γ.*

Then for a specific site, the rate matrix for this site is $\gamma_j \times Q$, where $j \in \{1, \ldots, k\}$. For different sites, different elements of γ are taken so that the rates are variable on diffferent sites. For example, if for the first 50 sites on a pair of homologous DNA sequences, the rate of substitution is slower than the next 50 sites on the same pair of sequences. Then we consider the rate matrix for the first 50 sites is $\gamma_1 \times Q$ and $\gamma_2 \times Q$ for the second 50 sites where $\gamma_1 \neq \gamma_2$. Then the probability for a nucleotide with state A to change into G is $\exp(\gamma_1 Qt)_{\{A \to G\}}$ for a site within the first 50 sites. The probability for a nucleotide with state A to change into G is $\exp(\gamma_2 Qt)_{\{A \to G\}}$ for a site within the second 50 sites.

Chapter 3

Likelihood Methods

3.1 Introduction

In this chapter, we introduce how to calculate the likelihood of a pair of DNA sequences under two situations. Identical rates are considered first and rate heterogeneity is introduced later. The computation of the likelihood for a pair of sequences serves as a basis for more complicated situations. After that, we explain how to calculate the likelihood of a given tree. However, in real cases, we need to make inference of the topology of the tree first. The details of how to infer the topology of the tree and the branch lengths of the tree are introduced by Felsenstein (2004).

3.2 Likelihood for a pair of DNA sequences

3.2.1 Without Rate Heterogeneity

Considering two homologous DNA sequences represented by $x^T = (x_1, x_2, \ldots, x_n)$ and $y^T = (y_1, y_2, ..., y_n)$, we first study how to calculate the likelihood for this pair of sequences. The parameters are $\theta = (Q, t)$, where Q is the instantaneous transition rate matrix and t is the branch length between this pair of sequences.

For a single site i on both sequences x and y, the likelihood is

$$L(\theta) = P(X_i = x_i)(e^{Qt})_{x_i \to y_i},$$

where $x_i, y_i \in \{A, G, T, C\}$, $P(X_i = x_i)$ is the base frequency $\pi = (\pi_A, \pi_C, \pi_G, \pi_T)$ which can be derived from the eigenvector of Q^T. If $x_i = A$, then $P(X_i = x_i) = \pi_A$.

For n sites on the sequences, under the assumption of independence across different sites, the likelihood is

$$L(\theta) = \prod_{i=1}^{n} P(X_i = x_i)(e^{Qt})_{x_i \to y_i}$$

The incomplete log likelihood for n sites on one sequence is

$$l(\theta) = \sum_{i=1}^{n} \log \left(P(X_i = x_i)(e^{Qt})_{x_i \to y_i} \right).$$

3.2.2 With Rate Heterogeneity

If different sites evolve at different rates, a rate factor γ is introduced. Recall that in Section 2.2.4 in Chapter 2, we have defined that the rate matrix for the ith site is $\gamma_i \times Q$, where γ_i is the rate factor for the ith site. In the literature, most authors assume the rate factor γ follows a continuous distribution like the gamma distribution or log normal distribution. The parameters are $\theta = (Q, t, \gamma)$. Assuming $g(\gamma)$ is the prior density function for γ, the likelihood for this pair of sequences is

$$L(\theta) = \prod_{i=1}^{n} \int_0^{\infty} P(X_i = x_i)(e^{\gamma_i Q t})_{x_i \to y_i} g(\gamma_i) d(\gamma_i). \tag{3.1}$$

In this case, we do not estimate the rate at each site, if a single rate is assumed for each site, then the number of unknowns increases quickly as the number of sites increases. As a result, as shown in Equation 3.1, we integrate out all the possible rates for each site and take it as the whole contribution of this site to the whole likelihood.

3.3 Likelihood for Trees

3.3.1 Without Rate Heterogeneity

Felsenstein (1981) has introduced the "pruning" method to economize the computation of the likelihood of the tree. In the field of computer science, this method is known under the name of "dynamic programming". The details of how to implement this method to a tree is covered in Chapter 16 of Felsenstein (2004). We will introduce this method briefly.

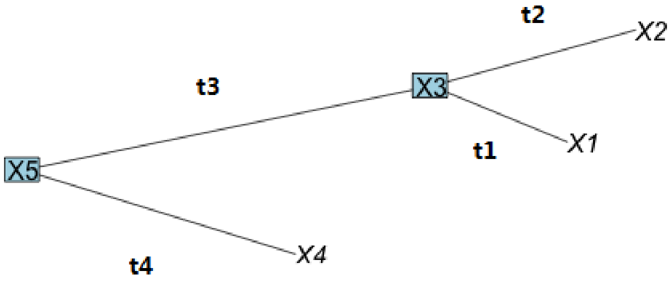

Figure 3.1: Tree Topology

We first illustrate how to calculate the likelihood for the tree in Figure 3.1 for only one site. Assuming for the ith site, the states of DNA at the three tips X_4, X_1 and X_2 are A, G and C respectively. From the tips to the root of the tree, the algorithm calculates the conditional likelihood of a subtree recursively given the state of this node in the current generation by combining the conditional likelihood of its immediate child in the left lineages and also the one in the right lineages of the same node.

Using the small tree, we illustrate how this algorithm works. The depth of the tree is two. Let $L^i_{x_k}(A)$ denote the probability of observing all the tips given the state of the node x_k has state A at the ith site. As assumed, at the tips, the state of X_2 in the ith site is C. Then $(L^i_{x_2}(A), L^i_{x_2}(C), L^i_{x_2}(G), L^i_{x_2}(T))$ $= (0,1,0,0)$, $(L^i_{x_2}(A), L^i_{x_2}(C), L^i_{x_2}(G), L^i_{x_2}(T)) = (0,0,1,0)$, $(L^i_{x_4}(A), L^i_{x_4}(C),$ $L^i_{x_4}(G), L^i_{x_4}(T)) = (1,0,0,0)$. For an internal node X_3, we illustrate how to calculate $L^i_{x_3}(A)$ as an example.

$$\begin{aligned} L^i_{x_3}(A) &= \left(\sum_x Prob(X_1 = x|A, t_1) L^i_{x_1}(x)\right) \\ &\quad \times \left(\sum_y Prob(X_2 = y|A, t_2) L^i_{x_2}(y)\right) \\ &= \left(\sum_x \exp(t_1 Q)_{A\to x} L^i_{x_1}(x)\right) \\ &\quad \times \left(\sum_y \exp(t_2 Q)_{A\to y} L^i_{x_1}(y)\right) \end{aligned}$$

As a result, $L^i_{x_3}(C), L^i_{x_3}(G), L^i_{x_3}(T))$ can be calculated in the same way.

$$\begin{aligned} L^i_{x_5}(r) &= \left(\sum_s Prob(X_3 = s|r, t_1) L^i_{x_3}(s)\right) \\ &\quad \times \left(\sum_q Prob(X_4 = q|r, t_4) L^i_{x_4}(q)\right) \\ &= \left(\sum_s \exp(t_3 Q)_{r\rightarrow s} L^i_{x_3}(s)\right) \\ &\quad \times \left(\sum_q \exp(t_4 Q)_{r\rightarrow q} L^i_{x_4}(q)\right) \end{aligned}$$

Assuming the stationary distribution for the four states is $\pi = (\pi_A, \pi_C, \pi_G, \pi_T)$, then the likelihood of this tree for the ith site is

$$L^i = \sum_r \pi_r L^i_{x_5}(r).$$

Denote the log-likelihood for the ith site as $l^i = \log(L^i)$. By assuming independence of different sites, the log-likelihood for this tree is $l = \sum_{i=1}^n l^i$.

3.3.2 With Rate Heterogeneity

However, it seems impractical that each site evolves at the same rate. Yang (1993) proposes the rate factor coming from a gamma distribution on each site. Below we provide how to calculate the likelihood of the previous tree in Figure 3.1 with the assumption that the rate follows the gamma distribution with mean one. Denote the prior density function for the rate γ as $g(\gamma)$. The difference of computing the likelihood when considering rate heterogeneity lies in its impossibility of "dynamic programming". Same as in Section 3.3.1, we explain how to calculate the likelihood of the tree in Figure 3.1 for only one site first. It is assumed the states of X_4, X_1 and X_2

are A, G and C respectively. The likelihood of the tree in Figure 3.1 for this site is

$$\sum_p \sum_q \pi_p \mathrm{Prob}(A|X_5 = p, t_4)\mathrm{Prob}(X_3 = q|p)\mathrm{Prob}(G|q, t_1)\mathrm{Prob}(C|q, t_2)$$
$$= \sum_p \sum_q \pi_p \int_0^\infty \pi_p g(\gamma) \exp(t_4 Q\gamma)_{p\to A} \exp(t_3 Q\gamma)_{p\to q} \exp(t_1 Q\gamma)_{q\to G}$$
$$\exp(t_2 Q\gamma)_{q\to C} d\gamma \tag{3.2}$$

From Equation 3.2, we can see that if the number of the internal nodes is w, the computation complexity is 4^w since "dynamic programming" can not be applied. The computation time of this method increases explosively as the number of species increases. As mentioned in the original paper, this method can only deal with tree topologies with no more than four species with a microcomputer.

In order to deal with the intense computation, Yang (1994) proposed the "discrete gamma distribution" to simplify the problem. By assuming equal probability of the rate in each category, Yang (1994) uses the mean or median in each category to represent all the rates in the same category. Assuming we use k categories of "discrete gamma distribution" to replace the continuous gamma distribution, where $(\gamma_1, \gamma_2, \ldots, \gamma_k)$ is the mean of each category. We use $L^i_{x_5}(r)$ to illustrate how to use the "discrete gamma distribution" to approximate the continuous case, where r is the state of X_5.

$$\begin{aligned}
L^i_{x_5}(r) &= \left(\sum_s Prob(X_3 = s|r, t_1)L^i_{x_3}(s)\right) \\
&\quad \times \left(\sum_q Prob(X_4 = q|r, t_4)L^i_{x_4}(q)\right) \\
&= \sum_{j=1}^{k} \frac{1}{k}\left(\sum_s \exp(\gamma_j t_3 Q)_{r\to s} L^i_{x_3}(s)\right) \\
&\quad \times \left(\sum_q \exp(\gamma_j t_4 Q)_{r\to q} L^i_{x_4}(q)\right)
\end{aligned}$$

As to how to calculate the mean of each category, readers can refer to the paper of Yang (1994) for details.

However, in both the continuous and the discrete cases, the author restricts the mean of the gamma distribution to one, then only the shape parameter α of the gamma distribution needs to be estimated. Since with the fixed number of species at the tips, a tree can have several different topologies. The process of maximizing the likelihood of a tree with a given topology is repeated for each of the possible topology until a maximum tree is found. By maximizing the likelihood of the tree over the branch lengths and α, an estimate of α can be obtained.

Both the continuous and discrete gamma distributions of the rate factor have their own limitations. Yang (1996) reviewed that by assuming the continuous gamma distribution of the rate factor, the algorithm is practical for no more than six sequences. By using the discrete gamma distribution, Yang (1994) has pointed out that $\hat{\alpha}$ can be very different based on different tree topologies where $\hat{\alpha}$ is the estimate of the shape parameter α of the gamma distribution. For example, $\hat{\alpha}$ tends to be larger given the maximum likelihood tree while $\hat{\alpha}$ seems the smallest from a star tree most of the times.

The difference gets larger especially when there are many species. As a result, we are motivated to relax the restriction of the specific distribution of γ because of the intense computation of the continuous gamma distribution and the overestimation or underestimation of the shape parameter α from the discrete gamma distribution given different tree structures. Finally, we propose a discrete distribution of the rate factor without any other restrictions which will be covered by the Chapter 4.

Chapter 4

Estimating the Rate Factor γ via an EM Algorithm

4.1 Introduction

In this chapter, we first introduce how to use an EM (Expectation Maximization) algorithm to estimate the rate factor and its distribution. We first introduce how to implement the method to a pair of DNA sequences. Then we explain how to estimate γ and its distribution for a tree.

4.2 Learn γ and f via an EM Algorithm for a pair of DNA sequences

4.2.1 The likelihood for a pair of DNA sequences

When having two DNA sequences represented by $x^T = (x_1, x_2, \ldots, x_n)$ and $y^T = (y_1, y_2, \ldots, y_n)$ which are the observed data, we will first study how to calculate the likelihood for this pair of sequences while the rate factor following a discrete distribution is introduced. Assuming that different nucleotide sites are independent within each DNA sequence, the rate factor is denoted as $\gamma^T = (\gamma_1, \gamma_2, \ldots, \gamma_k)$ which is used to scale the original rate matrix Q and a latent variable Z is introduced to imply which category the rate for a particular site belongs to. The latent variable Z is the missing data. The complete data includes the observed sequences x, y and z which is the realization of the latent variable Z. The prior distribution of Z is $f = (f_1, f_2, \ldots, f_k)$. The parameters to be estimated are $\theta = (\gamma, f)$ given the data from pairs of DNA sequences.

For a single site i on both sequences x and y, the likelihood is

$$L(\theta) = \sum_{j=1}^{k} P(Z_i = j)(e^{\gamma_j Qt})_{x_i \to y_i} P(X_i = x_i),$$

where $x_i, y_i \in \{A, G, T, C\}$, $i \in \{1, 2, \ldots, n\}$ and $P(X_i = x_i)$ is the base frequency which can be derived from the eigenvector of Q^T.

For n sites on the sequences, under the assumption of independence across different sites, the likelihood is

$$L(\theta) = \prod_{i=1}^{n} \sum_{j=1}^{k} P(Z_i = j)(e^{\gamma_j Qt})_{x_i \to y_i} P(X_i = x_i).$$

The incomplete log-likelihood for n sites on one sequence is

$$l(\theta) = \sum_{i=1}^{n} \log \left(\sum_{j=1}^{k} P(Z_i = j)(e^{\gamma_j Qt})_{x_i \to y_i} P(X_i = x_i) \right).$$

Since this incomplete log likelihood is hard to deal with, an EM algorithm is introduced. In order to utilize the EM algorithm, we need to calculate the complete log likelihood for this pair of DNA sequences first.

For the ith site on the DNA sequence, z_i represents which category the rate γ takes. The complete likelihood for a pair of sequences with n sites given γ is

$$L(\theta; x, y, z) = \prod_{i=1}^{n} \prod_{j=1}^{k} f_j^{1(z_i=j)} (e^{\gamma_j Qt})_{x_i \to y_i}^{1(z_i=j)} P(X_i = x_i).$$

The complete log-likelihood given the parameters γ will be

$$l(\theta; x, y, z) = \sum_{i=1}^{n} \sum_{j=1}^{k} 1(z_i = j) \log \left(f_j (e^{\gamma_j Qt})_{x_i \to y_i} P(X_i = x_i) \right).$$

Denote $P(X = x) = \pi_x$ as the prior probability, the complete log-likelihood will be

$$l(\theta; x, y, z) = \sum_{i=1}^{n} \sum_{j=1}^{k} 1(z_i = j) \log \left(f_j \pi_{x_i} (e^{\gamma_j Qt})_{x_i \to y_i} \right).$$

The expected complete log-likelihood is

$$E(l(\theta; x, y, Z)|X = x, Y = y) = \sum_{i=1}^{n} \sum_{j=1}^{k} E(1(Z_i = j)|X_i = x_i, Y_i = y_i)$$

$$\log(f_j \pi_{x_i} (e^{\gamma_j Qt})_{x_i \to y_i})$$

When we have N pairs of DNA sequences, the expected complete log-likelihood is

$$E(l(\theta; x, y, Z)|X = x, Y = y)$$
$$= \sum_{m=1}^{N} \sum_{i=1}^{n_m} \sum_{j=1}^{k} E(1(Z_{m,i} = j)|X_{m,i} = x_{m,i}, Y_{m,i} = y_{m,i}) \log(f_j \pi_{x_{m,i}} (e^{\gamma_j Qt})_{x_{m,i} \to y_{m,i}})$$

The number of sites in different pairs of sequences can be different which is denoted by n_m.

4.2.2 Learn γ and f via an EM Algorithm

In this section, we explain how to learn the parameters $\gamma = (\gamma_1, \ldots, \gamma_k)$ and update the posterior distribution $f = (f_1, \ldots, f_k)$ for the latent variable Z given the DNA sequence data. When there are more than one pair of sequences, it needs to sum over all the sites on all pairs of DNA sequences to obtain the expected complete log-likelihood. For simplicity, we assume there is one pair of DNA sequences here.

E step of EM algorithm

$$
\begin{aligned}
&E_{\theta^{t-1}}(l(\theta;x,y,Z)|X=x,Y=y)\\
&=\sum_{i=1}^{n}\sum_{j=1}^{k}E_{\theta^{t-1}}(1(Z_i=j)|X_i=x_i,Y_i=y_i)\log(f_j\pi_{x_i}(e^{\gamma_j Qt})_{x_i\to y_i})\\
&=\sum_{i=1}^{n}\sum_{j=1}^{k}P_{\theta^{t-1}}(Z_i=j|X_i=x_i,Y_i=y_i)\log(f_j\pi_{x_i}(e^{\gamma_j Qt})_{x_i\to y_i})\\
&=\sum_{i=1}^{n}\sum_{j=1}^{k}\frac{P_{\theta^{t-1}}(Z_i=j,X_i=x_i,Y_i=y_i)}{P_{\theta^{t-1}}(X_i=x_i,Y_i=y_i)}\log(f_j\pi_{x_i}(e^{\gamma_j Qt})_{x_i\to y_i})
\end{aligned}
$$

Assuming $\theta^{t-1}=(\gamma^{t-1},f^{t-1})$ has been learned from $(t-1)$th step, where $\gamma^{t-1}=(\gamma_1^{t-1},\ldots,\gamma_k^{t-1})$ and $f^{t-1}=(f_1^{t-1},\ldots,f_k^{t-1})$. Then in the tth step, both γ_j and f_j will be updated, where $j=1,2,\ldots,k$.

The exectation term $E(1(Z_i=j)|X_i=x_i,Y_i=y_i)$ is taken with respect to the old parameters $\gamma^{t-1}=(\gamma_1^{t-1},\ldots,\gamma_k^{t-1})$ and the observed data x_i,y_i. The goal of the E step is to compute $E(l(\theta;x,y,Z)|X=x,Y=y)$. In the M step, since there is no analytical solution of γ, $E(l(\theta;x,y,Z)|X=x,Y=y)$ is optimized with respect to γ

$$
\begin{aligned}
&E_{\theta^{t-1}}(l(\theta;x,y,Z)|X=x,Y=y)\\
&=\sum_{i=1}^{n}\sum_{j=1}^{k}\frac{e^{(\gamma_j^{t-1}Qt)}{}_{x_i\to y_i}f_j^{(t-1)}}{\sum_{j=1}^{k}e^{(\gamma_j^{t-1}Qt)}{}_{x_i\to y_i}f_j^{(t-1)}}\log(f_j\pi_{x_i}(e^{\gamma_j Qt})_{x_i\to y_i}). \qquad (4.1)
\end{aligned}
$$

M step of the EM algorithm

Updating f_j

The posterior distribution $f=(f_1,\ldots,f_k)$ of latent variable Z is updated in the following way in the tth step assuming $f^{(t-1)}$ and $\gamma^{(t-1)}$ have been obtained.

Denote

$$A_j^{t-1} = \frac{e^{(\gamma_j^{t-1}Qt)} f_j^{(t-1)}}{\sum_{j=1}^{k} e^{(\gamma_j^{t-1}Qt)} f_j^{(t-1)}} \tag{4.2}$$

as a 4×4 matrix, $j = 1, 2, \ldots, k$. The element in the ith row and kth column of A_j^{t-1} represents the posterior probability of $Z = j$ meaning that γ takes the jth category given the site changes from state i to state k of two homologous sequences from a common ancestor in the $(t-1)$th iteration.

There are four states A, G, T, C in the DNA sequence. As a result there are 16 kinds of transitions between these states including A→A, A→G, A →T and so on. For a pair of sequences with n sites, the number of each kind of transition will be counted and these numbers form a 4×4 matrix B. The element in matrix B records the number of this kind of transition corresponding to the element appearing in the Q matrix. That means the row and column order of {A, G, T, C} in B is the same as rate matrix Q and A_j^{t-1}. For example, if the element in the first row and second column in Q denotes the rate for a state to change from A to C, then the element in the first row and second column in B denotes the number of transitions from A to C for n sites on this pair of DNA sequences.

Let $B \cdot A_j^{t-1}$ denote the dot product of two matrix, where (l, m) denote the dimension of matrix $B \cdot A_j^{t-1}$. Then f_j, $j = 1, 2 \ldots, k$ is updated as for a pair of sequences with n sites as follows

$$f_j^{(t)} = \sum_{l,m} (B \cdot A_j^{t-1})_{l,m} / n. \tag{4.3}$$

The sum of all elements in $B \cdot A_j^{t-1}$ is the expected weighted number of sites for γ to take the jth category, $j = 1, \ldots, k$.

Updating γ

To update γ in the tth step, we choose γ such that a local maximum of Equation 4.1 can be obtained when the gradient of γ is a zero vector and the Hessian matrix of γ is negative definite.

In order to deal with getting the gradient of Equation 4.1 with respect to γ, we first introduce Theorem 1 to illustrate how to get the derivative of a particular element of matrix exponential.

Theorem 1.

$$\lim_{t_n \to t} \frac{(e^{t_n Q})_{i,j} - (e^{tQ})_{i,j}}{t_n - t} = (Qe^{tQ})_{i,j}$$

Theorem 1 states that if we get derivative with respect to t for the element in the ith row and jth column of the matrix (e^{tA}) is equivalent to getting derivative with respect to t for the whole matrix and then take the element in the ith row and jth column. This result is obtained by Wilcox (1967) and we provide a standard proof of Theorem 1 in Appendix A.

With Theorem 1, we can get the gradient and Hessian matrix with respect to γ.

Review in Equation 4.1 and Equation 4.2, we have that

$$\begin{aligned} &E_{\theta^{t-1}}(l(\theta; x, y, Z)|X = x, Y = y) \\ &= \sum_{i=1}^{n} \sum_{j=1}^{k} \frac{(e^{(\gamma_j^{t-1} Qt)})_{x_i \to y_i} f_j^{(t-1)}}{\sum_{j=1}^{k} (e^{(\gamma_j^{t-1} Qt)})_{x_i \to y_i} f_j^{(t-1)}} \log(f_j \pi_{x_i} (e^{\gamma_j Qt})_{x_i \to y_i}), \end{aligned}$$

$$A_j^{t-1} = \frac{e^{(\gamma_j^{t-1} Qt)} f_j^{(t-1)}}{\sum_{j=1}^{k} e^{(\gamma_j^{t-1} Qt)} f_j^{(t-1)}},$$

then,

$$E_{\theta_{t-1}}(l(\theta; x, y, Z)|X = x, Y = y) = \sum_{i=1}^{n} \sum_{j=1}^{k} (A_j^{t-1})_{x_i \to y_i} \log(f_j \pi_{x_i} (e^{\gamma_j Qt})_{x_i \to y_i}) \quad (4.4)$$

In order to find the local maximum of Equation 4.4, the gradient of Equation 4.4 with respect to γ is desired. Denote $E(l(\theta; x, y, Z)|X = x, Y = y)$ as $W(\gamma)$, where $\gamma = (\gamma_1, \gamma_2, \dots, \gamma_k)$ and getting the derivative of $W(\gamma)$ is only with respect to γ. Assume γ_m is one element of γ, $m = 1, 2, \dots, k$. Under the assumption that γ_i and γ_j are independent when $i \neq j$, it can be obtained that

$$\begin{aligned}\frac{\partial W(\gamma)}{\partial \gamma_m} &= \textstyle\sum_{i=1}^{n} (A_j^{t-1})_{x_i \to y_i} \frac{f_m (tQ \cdot \exp(\gamma_m tQ))_{x_i \to y_i}}{(e^{(\gamma_m Qt)})_{x_i \to y_i} f_m} \\ &= \textstyle\sum_{i=1}^{n} (A_j^{t-1})_{x_i \to y_i} \frac{(tQ \cdot \exp(\gamma_m tQ))_{x_i \to y_i}}{(e^{(\gamma_m Qt)})_{x_i \to y_i}}. \end{aligned} \tag{4.5}$$

Finally we have,

$$\frac{\partial W(\gamma)}{\partial \gamma} = \left(\frac{\partial W(\gamma)}{\partial \gamma_1}, \frac{\partial W(\gamma)}{\partial \gamma_2} \dots, \frac{\partial W(\gamma)}{\partial \gamma_k} \right)^T. \tag{4.6}$$

How to calculate each element of $\frac{\partial W(\gamma)}{\partial \gamma}$ is defined in Equation 4.5.

Next, the Hessian matrix of $W(\gamma)$ with respect to γ is deducted. Under the assumption that γ_i and γ_j are independent when $i \neq j$, the Hessian matrix is an diagonal matrix. For any $\gamma_m \in \gamma$, where $m \in 1, 2, \dots, k$, we have

$$\begin{aligned} \frac{\partial^2 W(\gamma)}{\partial^2 \gamma_m} &= \sum_{i=1}^{n} (A_j^{t-1})_{x_i \to y_i} \Bigg(\frac{\frac{\partial}{\partial \gamma_m} [tQ \exp(\gamma_m tQ)]_{x_i \to y_i} (\exp(\gamma_m tQ))_{x_i \to y_i}}{\exp(\gamma_m tQ)^2_{x_i \to y_i}} \\ &\quad - \frac{[tQ \cdot \exp(\gamma_m tQ)]_{x_i \to y_i} \frac{\partial}{\partial \gamma_m} (\exp(\gamma_m tQ))_{x_i \to y_i}}{\exp(\gamma_m tQ)^2_{x_i \to y_i}} \Bigg) \\ &= \sum_{i=1}^{n} (A_j^{t-1})_{x_i \to y_i} \Bigg(\frac{[tQ \cdot tQ \cdot \exp(\gamma_m tQ)]_{x_i \to y_i} \exp(\gamma_k tQ)_{x_i \to y_i}}{\exp(\gamma_m tQ)^2_{x_i \to y_i}} \\ &\quad - \frac{[tQ \cdot \exp(\gamma_m tQ)]_{x_i \to y_i} [tQ \cdot \exp(\gamma_m tQ)]_{x_i \to y_i}}{\exp(\gamma_m tQ)^2_{x_i \to y_i}} \Bigg). \end{aligned} \tag{4.7}$$

The Hessian matrix is an $k \times k$ diagonal matrix with the mth diagonal element defined in Equation 4.7.

Based on the gradient and Hessian matrix, γ satisfies that the gradient at this vector is zero and the Hessian matrix is negative positive. This can be obtained by using optim() function in R to minimize the negative log likelihood and specifying the gradient and Hessian matrix.

When there are multiple sequences, for the expected log likelihood, gradient and Hessian matrix, we just sum over all the sites on all pairs of sequences.

The iteration to update γ and f_j ends when γ^{t-1} and γ^t are quite close, for example, $||\gamma^t - \gamma^{t-1}|| \leq 10^{-5}$.

4.3 Likelihood for a tree with multiple DNA sequences

We have covered how to obtain the likelihood for a pair of DNA sequences with multiple sites. Now we concentrate on how to get the likelihood of a tree. We begin with the simplest tree with three DNA sequences. There are three tips and two internal nodes in this situation. The topology of the tree is shown in Figure 4.1.

In this tree, X_1, X_2 and X_4 are three DNA sequences at tips of the tree representing three different species. For any single site on X_1, X_2 or X_4, we have already known the state of this site. The state space is $\Omega = \{A, G, T, C\}$, X_3 and X_5 are annotated since they are internal nodes. For any single site on an internal node, it can take A, G, T or C which is not known to us. The notations of t_1, t_2, t_3 and t_4 represent the branch lengths in the tree.

First we show how to get the likelihood of this tree given the DNA sequence data at X_1, X_2 and X_4. Independence among different nucleotide

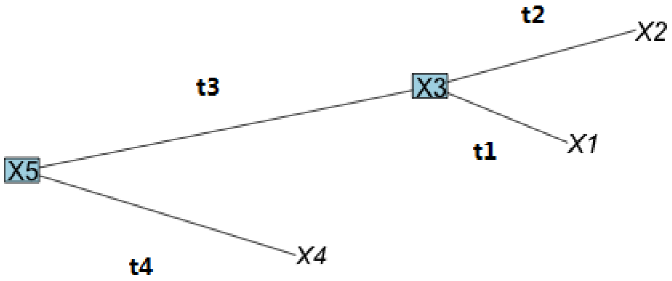

Figure 4.1: Tree Topology

sites is assumed within each DNA sequence. The rate factor $\gamma^T = (\gamma_1, \gamma_2, \ldots, \gamma_k)$ is defined the same as in the previous section. A latent variable Z is introduced to imply which category of the rate factor a particular site takes. The distribution of Z is $f = (f_1, f_2, \ldots, f_k)$. The stationary distribution for {A, G, T, C} is $\pi = (\pi_A, \pi_G, \pi_T, \pi_C)$. The parameters to be estimated are $\theta = (\gamma, f)$ given the DNA sequences from X_1, X_2, and X_4 at the tips.

For a single site i on both the ancestral and descendent sequences, the likelihood is

$$\begin{aligned} L(\theta) &= \sum_{z=1}^{k} \sum_{\substack{p\in\Omega \\ x_5=p}} \sum_{\substack{q\in\Omega \\ x_3=q}} f_z \pi_p \exp(\gamma_z Q t_4)_{p\to x_4} \cdot \exp(\gamma_z Q t_3)_{p\to q} \\ & \quad \exp(\gamma_z Q t_1)_{q\to x_1} \exp(\gamma_z Q t_2)_{q\to x_2}, \end{aligned}$$

where $\Omega = \{A, G, T, C\}$ and π_p is the stationary frequency which can be derived from the eigenvector of Q^T.

For n sites on the sequences, under the assumption of independence across different sites, the likelihood is

$$
\begin{aligned}
L(\theta) \quad = \quad & \prod_{i=1}^{n}\sum_{z=1}^{k}\sum_{\substack{p\in\Omega \\ x_5=p}}\sum_{\substack{q\in\Omega \\ x_3=q}} f_z\pi_p \exp(\gamma_z Qt_4)_{p\to x_4}\cdot\exp(\gamma_z Qt_3)_{p\to q} \\
& \exp(\gamma_z Qt_1)_{q\to x_1}\exp(\gamma_z Qt_2)_{q\to x_2}
\end{aligned}
$$

The incomplete log-likelihood for n sites of this tree is

$$
\begin{aligned}
& l(\theta; x_1, x_2, x_4) \\
& = \sum_{i=1}^{n}\log\Bigg(\sum_{z=1}^{k}\sum_{\substack{p\in\Omega \\ x_{5i}=p}}\sum_{\substack{q\in\Omega \\ x_{3i}=q}} f_z\pi_p \exp(\gamma_z Qt_4)_{p\to x_{4i}}\cdot\exp(\gamma_z Qt_3)_{p\to q} \\
& \exp(\gamma_z Qt_1)_{q\to x_{1i}}\exp(\gamma_z Qt_2)_{q\to x_{2i}}\Bigg)
\end{aligned}
$$

This incomplete log-likelihood is hard to deal with since we need to marginalize all the internal nodes and latent variables, an EM algorithm is introduced for trees with multiple DNA sequences as well.

4.3.1 Complete Likelihood for Multiple DNA sequences

The complete likelihood for this tree with n nucleotide sites is

$$
\begin{aligned}
& L(\gamma, z; x_1, x_2, x_4) \\
& = \prod_{i=1}^{n}\prod_{j=1}^{k}\prod_{x_{5i}\in\Omega}\prod_{x_{3i}\in\Omega}\Bigg(f_j\pi_p \exp(\gamma_j Qt_4)_{x_{5i}\to x_{4i}}\cdot\exp(\gamma_j Qt_3)_{x_{5i}\to x_{3i}} \\
& \exp(\gamma_j Qt_1)_{x_{3i}\to x_{1i}}\exp(\gamma_j Qt_2)_{x_{3i}\to x_{2i}}\Bigg)^{(1(z_i=j,X_{5i}=p,X_{3i}=q))} \qquad (4.8)
\end{aligned}
$$

Then the complete log-likelihood is

$$
\begin{aligned}
& l(\gamma, z; x_1, x_2, x_4) \\
& = \sum_{i=1}^{n} \sum_{j=1}^{k} \sum_{x_{5i} \in \Omega} \sum_{x_{3i} \in \Omega} 1(z_i = j, X_{5i} = p, X_{3i} = q) \times \log(f_j \pi_p \\
& \exp(\gamma_j Q t_4)_{p \to x_{4i}} \cdot \exp(\gamma_j Q t_3)_{p \to q} \exp(\gamma_j Q t_1)_{p \to x_{1i}} \\
& \exp(\gamma_j Q t_2)_{q \to x_{2i}}) \qquad (4.9)
\end{aligned}
$$

Denote $E(l(\gamma, Z; x_1, x_2, x_4)|X_1 = x_1, X_2 = x_2, X_4 = x_2)$ as $W(\gamma)$, then the expected complete log-likelihood is

$$
\begin{aligned}
& W(\gamma) \\
& = \sum_{i=1}^{n} \sum_{j=1}^{k} \sum_{x_{5i} \in \Omega} \sum_{x_{3i} \in \Omega} E(1(Z_i = j, X_{5i} = p, X_{3i} = q)|X_{1i} = x_{1i}, X_{2i} = x_{2i}, X_{4i} = x_{4i}) \\
& \times \log(f_j \pi_p \exp(\gamma_j Q t_4)_{p \to x_{4i}} \exp(\gamma_j Q t_3)_{p \to q} \exp(\gamma_j Q t_1)_{q \to x_{1i}} \exp(\gamma_j Q t_2)_{q \to x_{2i}}) \\
& \qquad (4.10)
\end{aligned}
$$

In order to simplify the calculation of function $W(\gamma)$, we separate each term of

$$\log(f_j \pi_p \exp(\gamma_j Q t_4)_{p \to x_{4i}} \exp(\gamma_j Q t_3)_{p \to q} \exp(\gamma_j Q t_1)_{q \to x_{1i}} \exp(\gamma_j Q t_2)_{q \to x_{2i}})$$

into
$\log(f_j), \log(\pi_p), \log(\exp(\gamma_j Q t_4)_{p \to x_{4i}}), \log(\exp(\gamma_j Q t_3)_{p \to q}), \log(\exp(\gamma_j Q t_1)_{q \to x_{1i}}),$
$\log(\exp(\gamma_j Q t_2)_{q \to x_{2i}}))$.

The following proposition reveals the details of how to calculate Equation 4.10 efficiently.

Proposition 1. *Assume there are n latent variables $X_1, X_2, \ldots, X_n$, $v_1, v_2, \ldots, v_n$ are functions that v_1 only depends on X_1, v_2 only depends on X_2, ... and v_n only depends on X_n. The denote a subset $\{1, 2, \ldots, n_1\}$ of $\{1, 2, \ldots, n\}$ as S. Then*

$$\begin{aligned}
&\sum_{x_1}\sum_{x_2}\ldots\sum_{x_n} E\left(1(X_1 = x_1, X_2 = x_2, \ldots, X_n = x_n)\right)\log(v_1 v_2 \ldots v_{n1}) \\
&= \sum_{S} E\left(1(X_1 = x_1, X_2 = x_2, \ldots, X_n = x_{n1})\right)\log(v_1 v_2 \ldots v_{n1}) \qquad (4.11)
\end{aligned}$$

Proof. The proof of Proposition 1 is simple. Denote $S_0 = \{1, 2, \ldots, n\}$. Since $S = \{1, 2, \ldots, n_1\}$, $S_1 = S_0 \backslash S$, then

$$\begin{aligned}
&\sum_{x_1}\sum_{x_2}\ldots\sum_{x_n} E\left(1(X_1 = x_1, X_2 = x_2, \ldots, X_n = x_n)\right)\log(v_1 v_2 \ldots v_{n_1}) \\
&= \sum_{S}\sum_{S_1} E\left(1(X_1 = x_1, X_2 = x_2, \ldots, X_n = x_n)\right)\log(v_1 v_2 \ldots v_{n_1}) \\
&= \sum_{S} E\left(1(X_1 = x_1, X_2 = x_2, \ldots, X_{n1} = x_{n_1})\right)\log(v_1 v_2 \ldots v_{n_1})
\end{aligned}$$

□

In order to calculate $W(\gamma)$, we separate the expected complete log-likelihood into several parts by partition the term in the product inside the log function and utilize Proposition 1 to calculate each of the following term and then sum them together.

Some of the latent variables can be summed over in the previous term to simplify the calculation.

$$\sum_{i=1}^{n}\sum_{j=1}^{k}\sum_{\substack{p\in\Omega\\ x_{5i}=p}}\sum_{\substack{q\in\Omega\\ x_{3i}=q}} E(1(Z_i=j,X_{5i}=p,X_{3i}=q)|X_{1i}=x_{1i},X_{2i}=x_{2i},X_{4i}=x_{4i})$$

$$\times \log(f_j)$$

$$=\sum_{j=1}^{k}\sum_{i=1}^{n} E(1(Z_i=j)|X_{1i}=x_{1i},X_{2i}=x_{2i},X_{4i}=x_{4i})\times\log(f_j) \tag{4.12}$$

$$\sum_{i=1}^{n}\sum_{j=1}^{k}\sum_{\substack{p\in\Omega\\ x_{5i}=p}}\sum_{\substack{q\in\Omega\\ x_{3i}=q}} E(1(Z_i=j,X_{5i}=p,X_{3i}=q)|X_{1i}=x_{1i},X_{2i}=x_{2i},X_{4i}=x_{4i})$$

$$\times \log(\pi_p)$$

$$=\sum_{j=1}^{k}\sum_{i=1}^{n}\sum_{\substack{p\in\Omega\\ x_{5i}=p}} E(1(Z_i=j,X_{5i}=p)|X_{1i}=x_{1i},X_{2i}=x_{2i},X_{4i}=x_{4i})\log(\pi_p) \tag{4.13}$$

$$\sum_{i=1}^{n}\sum_{j=1}^{k}\sum_{\substack{p\in\Omega\\ x_{5i}=p}}\sum_{\substack{q\in\Omega\\ x_{3i}=q}} E(1(Z_i=j,X_{5i}=p,X_{3i}=q)|X_{1i}=x_{1i},X_{2i}=x_{2i},X_{4i}=x_{4i})$$

$$\times \log(\exp(\gamma_j Q t_4)_{p\to x_{4i}})$$

$$=\sum_{j=1}^{k}\sum_{i=1}^{n}\sum_{\substack{p\in\Omega\\ x_{5i}=p}} E(1(Z_i=j,X_{5i}=p,X_{4i}=x_{4i})|X_{1i}=x_{1i},X_{2i}=x_{2i},X_{4i}=x_{4i})$$

$$\log(\exp(\gamma_j Q t_4)_{p\to x_{4i}}) \tag{4.14}$$

$$\sum_{i=1}^{n}\sum_{j=1}^{k}\sum_{\substack{p\in\Omega\\ x_{5i}=p}}\sum_{\substack{q\in\Omega\\ x_{3i}=q}} E(1(Z_i=j,X_{5i}=p,X_{3i}=q)|X_{1i}=x_{1i},X_{2i}=x_{2i},X_{4i}=x_{4i})$$

$$\times\log(\exp(\gamma_jQt_3)_{p\to q})$$

$$=\sum_{j=1}^{k}\sum_{i=1}^{n}\sum_{\substack{p\in\Omega\\ x_{5i}=p}}\sum_{\substack{q\in\Omega\\ x_{3i}=q}} E(1(Z_i=j,X_{5i}=p,X_{3i}=q)|X_{1i}=x_{1i},X_{2i}=x_{2i},X_{4i}=x_{4i})$$

$$\log(\exp(\gamma_jQt_3)_{p\to q}) \tag{4.15}$$

$$\sum_{i=1}^{n}\sum_{j=1}^{k}\sum_{\substack{p\in\Omega\\ x_{5i}=p}}\sum_{\substack{q\in\Omega\\ x_{3i}=q}} E(1(Z_i=j,X_{5i}=p,X_{3i}=q)|X_{1i}=x_{1i},X_{2i}=x_{2i},X_{4i}=x_{4i})$$

$$\times\log(\exp(\gamma_jQt_3)_{q\to x_{1i}})$$

$$=\sum_{j=1}^{k}\sum_{i=1}^{n}\sum_{\substack{q\in\Omega\\ x_{3i}=q}} E(1(Z_i=j,X_{3i}=q,X_{1i}=x_{1i})|X_{1i}=x_{1i},X_{2i}=x_{2i},X_{4i}=x_{4i})$$

$$\log(\exp(\gamma_jQt_1)_{q\to x_{1i}}) \tag{4.16}$$

$$\sum_{i=1}^{n}\sum_{j=1}^{k}\sum_{\substack{p\in\Omega\\ x_{5i}=p}}\sum_{\substack{q\in\Omega\\ x_{3i}=q}} E(1(Z_i=j,X_{5i}=p,X_{3i}=q)|X_{1i}=x_{1i},X_{2i}=x_{2i},X_{4i}=x_{4i})$$

$$\times\log(\exp(\gamma_jQt_3)_{q\to x_{2i}})$$

$$=\sum_{j=1}^{k}\sum_{i=1}^{n}\sum_{\substack{q\in\Omega\\ x_{3i}=q}} E(1(Z_i=j,X_{3i}=q,X_{2i}=x_{2i})|X_{1i}=x_{1i},X_{2i}=x_{2i},X_{4i}=x_{4i})$$

$$\log(\exp(\gamma_{Z_i}Qt_2)_{q\to x_{2i}} \tag{4.17}$$

Then $W(\gamma)$ is the sum of Equation 4.12, Equation 4.13, … and Equation 4.17.

Equation 4.17 reflects that the indicator function only depends on the states of the ith site of the ancestral and its direct descendant sequences. For example,

$$\sum_{i=1}^{n}\sum_{j=1}^{k}\sum_{\substack{p\in\Omega\\x_{5i}=p}}\sum_{\substack{q\in\Omega\\x_{3i}=q}} E(1(Z_i=j,X_{5i}=p,X_{3i}=q)|X_{1i}=x_{1i},X_{2i}=x_{2i},X_{4i}=x_{4i})$$

$$\log(\exp(\gamma_j Qt_2)_{q\to x_{2i}})$$

$$=\sum_{j=1}^{k}\sum_{i=1}^{n}\sum_{\substack{q\in\Omega\\x_{3i}=q}} E(1(Z_i=j,X_{3i}=q,X_{2i}=x_{2i})|X_{1i}=x_{1i},X_{2i}=x_{2i},X_{4i}=x_{4i})$$

$$\log(\exp(\gamma_j Qt_2)_{q\to x_{2i}})$$

When we are interested in calculating $\sum_{j=1}^{k}\sum_{i=1}^{n}\sum_{\substack{p\in\Omega\\x_{5i}=p}}\sum_{\substack{q\in\Omega\\x_{3i}=q}} E(1(Z_i=j,X_{5i}=p,X_{3i}=q)|X_{1i}=x_{1i},X_{2i}=x_{2i},X_{4i}=x_{4i})\log(\exp(\gamma_j Qt_2)_{q\to x_{2i}})$, where X_3 is the ancestral sequence and X_2 is X_3's descendant sequence. The states of internal nodes X_5 can be summed over so that X_5 does not appear in $\sum_{j=1}^{k}\sum_{i=1}^{n}\sum_{\substack{q\in\Omega\\x_{3i}=q}} E(1(Z_i=j,X_{3i}=q,X_{2i}=x_{2i})|X_{1i}=x_{1i},X_{2i}=x_{2i},X_{4i}=x_{4i})\log(\exp(\gamma_j Qt_2)_{q\to x_{2i}})$.

After separating the expected complete log-likelihood into several parts by partition each term in the product inside the log function, the indicator function only depends on a pair of ancestral and descendant node of interest which are X_{3i} and X_{2i} since all unknown states of other internal nodes can be summed over. Moreover, $1(Z_i=j,X_{3i}=q,X_{2i}=x_{2i})$ depends only on Z_i,X_{3i},X_{2i} which appears in the term that multiplies it which is $\log(\exp(\gamma_j Qt_2)_{x_{3i}\to x_{2i}})$.

Then to calculate $\sum_{j=1}^{k}\sum_{i=1}^{n}\sum_{p\in\Omega}\sum_{q\in\Omega} E(1(Z_i=j,X_{5i}=p,X_{3i}=q)|X_{1i}=x_{1i},X_{2i}=x_{2i},X_{4i}=x_{4i})\log(\exp(\gamma_j Qt_3)_{p\to q})$, we define a matrix E and a matrix L for computation convenience because we would like to denote this term as a sum of the dot product of each row in E and the corresponding row in L.

$$E = \begin{pmatrix} \sum_{i=1}^n E(1(Z_i = 1, x_{ai} = A, x_{di} = A)|data) & \dots & \sum_{i=1}^n E(1(Z_i = 1, x_{ai} = G, x_{di} = G)|data) \\ \vdots & \vdots & \vdots \\ \sum_{i=1}^n E(1(Z_i = k, x_{ai} = A, x_{di} = A)|data) & \dots & \sum_{i=1}^n E(1(Z_i = k, x_{ai} = G, x_{di} = G)|data) \end{pmatrix},$$

where x_{ai} denotes the ith site on the ancestral sequence and x_{di} denotes the ith site on the descendant sequence. We can just arrange

$$\log(\exp(\gamma_{z_i} Q t_3))$$

into a $k \times 16$ matrix L.

$$L = \begin{pmatrix} \log(\exp(\gamma_1 Q t)_{A \to A}) & \dots & \log(\exp(\gamma_1 Q t)_{G \to G}) \\ \vdots & \ddots & \vdots \\ \log(\exp(\gamma_k Q t)_{A \to A}) & \dots & \log(\exp(\gamma_k Q t)_{G \to G}) \end{pmatrix}$$

As a result, to calculate the expected complete likelihood of the tree in Figure 1, we can divide it into three parts. The first part is

$$\sum_{j=1}^{k} \sum_{i=1}^{n} E(1(Z_i = j)|x_{1i}, x_{2i}, x_{4i}) \log(f_j),$$

which can be obtained by summing over all the internal nodes.

The second part is

$$\sum_{j=1}^{k} \sum_{i=1}^{n} \sum_{p \in \Omega} E(1(Z_i = j, X_{5i} = p)|x_{1i}, x_{2i}, x_{4i}) \log(\pi_p).$$

To calculate it, we need to know the assumed location of the root of the tree and which is the root's direct child. By summing over the possible states of the descendant node, we can obtain this term.

In the third part, we compute the expected complete log-likelihood between each pair of ancestral and its direct desendant sequences and then sum them together. Now we introduce how to learn the parameters of γ and its distribution f.

When we perform the real data analysis, we add a penalty term which is the sum of γ^2 to the complete log-likelihood. The reason is that if any category of γ is so large then $\exp(\gamma Qt)$ will go to the stationary distribution. In that case, even if γ increases, the likelihood will not increase dramatically. As a result, we add the penalty term to constrain it.

4.4 Expectation Maximization to learn γ and f for Trees

According to Equation 4.10, we have the expression of how to calculate the expected complete log-likelihood.

4.4.1 E step

$$
\begin{aligned}
&E(l(\gamma, Z; x_1, x_2, x_4)|X_1 = x_1, X_2 = x_2, X_4 = x_4) \\
&= \sum_{i=1}^{n} \sum_{j=1}^{k} \sum_{\substack{p\in\Omega \\ x_{5i}=p}} \sum_{\substack{q\in\Omega \\ x_{3i}=q}} E(1(Z_i = j, X_{5i} = p, X_{3i} = q)|X_{1i} = x_{1i}, X_{2i} = x_{2i}, X_{4i} = x_{4i}) \\
&\times \log(f_j \pi_p \exp(\gamma_j Q t_4)_{p\to x_{4i}} \exp(\gamma_j Q t_3)_{p\to q} \exp(\gamma_j Q t_1)_{q\to x_{1i}} \exp(\gamma_j Q t_2)_{q\to x_{2i}})
\end{aligned}
\tag{4.18}
$$

$$E(1(Z_i = j, X_{5i} = p, X_{3i} = q)|X_{1i} = x_{1i}, X_{2i} = x_{2i}, X_{4i} = x_{4i}))$$
$$= \frac{P(Z_i = j, X_{5i} = p, X_{3i} = q, X_{1i} = x_{1i}, X_{2i} = x_{2i}, X_{4i} = x_{4i})}{\sum_{j=1}^{k}\sum_{p\in\Omega}\sum_{q\in\Omega} P(Z_i = j, X_{5i} = p, X_{3i} = q, X_{1i} = x_{1i}, X_{2i} = x_{2i}, X_{4i} = x_{4i})} \tag{4.19}$$

$$P(Z_i = j, X_{5i} = p, X_{3i} = q, X_{1i} = x_{1i}, X_{2i} = x_{2i}, X_{4i} = x_{4i})$$
$$= f_j \pi_p \exp(\gamma_j Q t_4)_{p\to t_4} \exp(\gamma_j Q t_3)_{p\to q} \exp(\gamma_j Q t_1)_{q\to x_1} \exp(\gamma_j Q t_2)_{q\to x_2} \tag{4.20}$$

$$E_{\theta^{t-1}}(1(Z_i = j, X_{5i} = p, X_{3i} = q)|X_{1i} = x_{1i}, X_{2i} = x_{2i}, X_{4i} = x_{4i})$$
$$= \frac{P_{\theta^{t-1}}(Z_i = j, X_{5i} = p, X_{3i} = q, X_{1i} = x_{1i}, X_{2i} = x_{2i}, X_{4i} = x_{4i})}{\sum_{j=1}^{k}\sum_{p\in\Omega}\sum_{q\in\Omega} P_{\theta^{t-1}}(Z_i = j, X_{5i} = p, X_{3i} = q, X_{1i} = x_{1i}, X_{2i} = x_{2i}, X_{4i} = x_{4i})}$$

where

$$P_{\theta^{t-1}}(Z_i = j, X_{5i} = p, X_{3i} = q, X_{1i} = x_{1i}, X_{2i} = x_{2i}, X_{4i} = x_{4i})$$
$$= f_j^{t-1} \pi_p \exp(\gamma_j^{t-1} Q t_4)_{p\to t_4} \exp(\gamma_j^{t-1} Q t_3)_{p\to q} \exp(\gamma_j^{t-1} Q t_1)_{q\to x_1} \exp(\gamma_j^{t-1} Q t_2)_{q\to x_2}$$

$$E_{\theta^{t-1}}(l(\gamma, Z; x_1, x_2, x_4)|X_1 = x_1, X_2 = x_2, X_4 = x_4)$$
$$= \sum_{i=1}^{n}\sum_{j=1}^{k}\sum_{\substack{p\in\Omega \\ x_{5i}=p}}\sum_{\substack{q\in\Omega \\ x_{3i}=q}} \frac{P_{\theta^{t-1}}(Z_i = j, X_{5i} = p, X_{3i} = q, X_1 = x_1, X_2 = x_2, X_4 = x_4)}{\sum_{j=1}^{k}\sum_{p\in\Omega}\sum_{q\in\Omega} P_{\theta^{t-1}}(Z_i = j, X_{5i} = p, X_{3i} = q, X_1 = x_1, X_2 = x_2, X_4 = x_4)}$$
$$\log(f_j \pi_p \exp(\gamma_j Q t_4)_{p\to t_4} \exp(\gamma_j Q t_3)_{p\to q} \exp(\gamma_j Q t_1)_{q\to x_1} \exp(\gamma_j Q t_2)_{q\to x_2}) \tag{4.21}$$

4.4.2 M step

Updating γ

In the M step, $E_{\theta^{t-1}}(l(\gamma; x_1, x_2, x_4, Z)|X_1, X_2, X_4)$ is optimized with respect to γ. Both γ and f_j are updated.

To learn parameters γ in the tth iteration, a local maximum with respect to γ can be obtained when the gradient of γ is a zero vector and the Hessian matrix of γ is negative definite. In order to get the gradient of γ, we need to know how to get derivatives with respect to γ_j in the following term:

$$\exp(\gamma_j Qt_4)_{p\to x_4}\exp(\gamma_j Qt_3)_{p\to q}\exp(\gamma_j Qt_1)_{q\to x_1}\exp(\gamma_j Qt_2)_{q\to x_2}.$$

To achieve that, we can prove the following theorem

Theorem 2.

$$\begin{aligned}
&\frac{\partial}{\partial\gamma}\left(\exp(\gamma Q_1)_{m\to n}\exp(\gamma Q_2)_{a\to b}\right)\\
&=[\frac{\partial}{\partial\gamma}(\exp(\gamma Q_1))]_{m\to n}\exp(\gamma Q_2)_{a\to b}+\exp(\gamma Q_1)_{m\to n}[\frac{\partial}{\partial\gamma}(\exp(\gamma Q_2))]_{a\to b}\\
&=(Q_1\exp(\gamma Q_1))_{m\to n}\exp(\gamma Q_2)_{a\to b}+\exp(\gamma Q_1)(Q_2\exp(\gamma Q_2))_{a\to b}
\end{aligned}$$

We provide a standard proof of this theorem in the Appendix.

In order to get the gradient with respect to γ_j, we first define

$$\begin{aligned}
&A_{i,j,p,q}\\
&=E(1(Z_i=j,X_{5i}=p,X_{3i}=q)|X_{1i},X_{2i},X_{4i})\\
&=\frac{P(Z_i=j,X_{5i}=p,X_{3i}=q,X_{1i}=x_{1i},X_{2i}=x_{2i},X_{4i}=x_{4i})}{\sum_{j=1}^{k}\sum_{p\in\Omega}\sum_{q\in\Omega}P(Z_i=j,X_{5i}=p,X_{3i}=q,X_{1i}=x_{1i},X_{2i}=x_{2i},X_{4i}=x_{4i})}.
\end{aligned} \tag{4.22}$$

$$E(l(\gamma, Z; x_1, x_2, x_4)|X_1 = x_1, X_2 = x_2, X_4 = x_4)$$

$$= \sum_{i=1}^{n} \sum_{j=1}^{k} \sum_{\substack{p \in \Omega \\ x_{5i}=p}} \sum_{\substack{q \in \Omega \\ x_{3i}=q}} E(1(Z_i = j, X_{5i} = p, X_{3i} = q)|X_{1i} = x_{1i}, X_{2i} = x_{2i}, X_{4i} = x_{4i})$$

$$\times \log(f_j \pi_p \exp(\gamma_j Q t_4)_{p \to x_{4i}} \exp(\gamma_j Q t_3)_{p \to x_{3i}} \exp(\gamma_j Q t_1)_{q \to x_{1i}} \exp(\gamma_j Q t_2)_{q \to x_{2i}})$$

$$= \sum_{i=1}^{n} \sum_{j=1}^{k} \sum_{p \in \Omega} \sum_{q \in \Omega} A_{i,j,p,q} \times \log(f_j \pi_p \exp(\gamma_j Q t_4)_{p \to x_{4i}} \exp(\gamma_j Q t_3)_{q \to q} \exp(\gamma_j Q t_1)_{q \to x_{1i}}$$

$$\exp(\gamma_j Q t_2)_{q \to x_{2i}})$$

$$E_{\theta^{t-1}}(l(\gamma, Z; x_1, x_2, x_4)|X_1 = x_1, X_2 = x_2, X_4 = x_4)$$

$$= \sum_{i=1}^{n} \sum_{j=1}^{k} \sum_{\substack{p \in \Omega \\ x_{5i}=p}} \sum_{\substack{q \in \Omega \\ x_{3i}=q}} E_{\theta^{t-1}}(1(Z_i = j, X_{5i} = p, X_{3i} = q)|X_{1i} = x_{1i}, X_{2i} = x_{2i}, X_{4i} = x_{4i})$$

$$\times \log(f_j \pi_q \exp(\gamma_j Q t_4)_{p \to x_{4i}} \exp(\gamma_j Q t_3)_{p \to q} \exp(\gamma_j Q t_1)_{q \to x_{1i}} \exp(\gamma_j Q t_2)_{q \to x_{2i}})$$

$$= \sum_{i=1}^{n} \sum_{j=1}^{k} \sum_{\substack{p \in \Omega \\ x_{5i}=p}} \sum_{\substack{q \in \Omega \\ x_{3i}=q}} A_{i,j,p,q}^{(t-1)} \times \log(f_j \pi_p \exp(\gamma_j Q t_4)_{p \to x_{4i}} \exp(\gamma_j Q t_3)_{p \to q} \exp(\gamma_j Q t_1)_{q \to x_{1i}}$$

$$\exp(\gamma_j Q t_2)_{q \to x_{2i}})$$

When we get derivatives with γ, denote $W_{\theta^{t-1}}(\gamma)$ as $E_{\theta^{t-1}}(l(\gamma; x_1, x_2, x_4)|X_1 =$

$x_1, X_2 = x_2, X_4 = x_4)$.

$$\begin{aligned}
\frac{\partial W_{\theta^{t-1}}(\gamma)}{\partial \gamma_j} &= \frac{\partial}{\partial \gamma_j} \sum_{i=1}^{n} \sum_{j=1}^{k} \sum_{\substack{p \in \Omega \\ x_{5i}=p}} \sum_{\substack{q \in \Omega \\ x_{3i}=q}} A_{j,p,q}^{(t-1)} \times \log(f_j \pi_p \exp(\gamma_j Q t_4)_{p \to x_{4i}} \\
& \exp(\gamma_j Q t_3)_{p \to q} \exp(\gamma_j Q t_1)_{q \to x_{1i}} \exp(\gamma_j Q t_2)_{q \to x_{2i}}) \\
&= \sum_{i=1}^{n} \sum_{\substack{p \in \Omega \\ x_{5i}=p}} \sum_{\substack{q \in \Omega \\ x_{3i}=q}} A_{i,j,p,q}^{(t-1)} \times \frac{\partial}{\partial \gamma_j} \log(f_j \pi_p \exp(\gamma_j Q t_4)_{p \to x_{4i}} \\
& \exp(\gamma_j Q t_3)_{p \to q} \exp(\gamma_j Q t_1)_{q \to x_{1i}} \exp(\gamma_j Q t_2)_{q \to x_{2i}}) \\
&= \sum_{i=1}^{n} \sum_{p \in \Omega} \sum_{q \in \Omega} A_{i,j,p,q}^{(t-1)} \Bigg(\frac{t_4 Q \exp(\gamma_j Q t_4)_{p \to x_4}}{\exp(\gamma_j Q t_4)_{p \to x_4}} + \frac{t_3 Q \exp(\gamma_j Q t_3)_{p \to q}}{\exp(\gamma_j Q t_4)_{p \to q}} \\
& + \frac{t_1 Q \exp(\gamma_j Q t_1)_{q \to x_1}}{\exp(\gamma_j Q t_1)_{q \to x_1}} + \frac{t_2 Q \exp(\gamma_j Q t_2)_{q \to x_2}}{\exp(\gamma_j Q t_2)_{q \to x_2}} \Bigg) \qquad (4.23)
\end{aligned}$$

Recall that for a pair of sequence with n nucleotide sites, $x_i, y_i \in \{A, G, T, C\}$

$$\frac{\partial W_{\theta^{t-1}}(\gamma)}{\partial \gamma_j} = \textstyle\sum_{i=1}^{n} (A_j^{t-1})_{x_i \to y_i} \frac{(tQ \cdot \exp(\gamma_j t Q))_{x_i \to y_i}}{(\exp(\gamma_j Q t))_{x_i \to y_i}}. \qquad (4.24)$$

When we have a tree, $\frac{\partial W_{\theta^{t-1}}(\gamma)}{\partial \gamma_j}$ is a sum over all pairs of the ancestor and its direct child sequences over all sites multiplying its corresponding posterior distribution $A_{i,j,p,q}^{(t-1)}$ from the E step.

Chapter 5

Identifiability of the Model

5.1 Introduction

In the context of phylogenetics, a model is non-identifiable if different set of parameters including tree topologies, branch lengths and evolutionary parameters can produce the same likelihood.

Section 5.2 is a review of the previous work on the identifiability of models with the rate factor following the gamma distribution. We discuss the relationship of the previous work and our work.

We illustrate how the identifiability of $\theta = (\gamma, f)$ is transformed into the problem of determining the uniqueness of the solution to a set of nonlinear equations in Section 5.3. The number of equations is equal to the number of different eigenvalues of Q. In Section 5.4, we prove that if the rate matrix comes from the F81 family, then the model is unidentifiable which carries over to the case where the rate factor follows a gamma distribution. In Section 5.5, by applying Wu and Susko (2010)'s result, we prove the non-identifiability of γ and f under certain conditions. Finally, we have shown some simulation results to provide empirical support of the non-identifiability of HKY85 model.

5.2 Review

Bryant, Galtier, and Poursat (2005) have reviewed both the identifiability of transition matrices for independent and identically distributed sites and models with rate variation by introducing a rate factor γ. The definition of the rate factor can be referred to Section 2.2.4. They review that if different sites evolve at the same rate, pairwise comparisons of sequences are not able to reconstruct the transition matrices and that the distribution of at least triples of sites is sufficient to reconstruct the transition matrices. Moreover, for models allowing rate variations, the topology and the transition rate matrices are identifiable under certain conditions. They can be identified when the distribution of γ is completely known. Assuming the rate factor follows a gamma distribution, Steel (2009) uses the F81 Model and discovers that the shape parameter of the gamma distribution and the topology of the tree are unidentifiable. Different shape parameters and tree topologies can produce the same pairwise distribution between all pairs of taxa. However, this is not a general case. Allman et al. (2008) proves that the four-state GTR + Γ model is identifiable given the joint distribution of at least triples of taxa by assuming the rate factor follows the gamma distribution. The reason why gamma distribution is favoured is its simple form of the inverse of moment generating function. It is still of interest whether pairwise distributions are sufficient to discover the identifiability of the model. Wu and Susko (2010) proves that several different conditions when the parameters can be identified. As long as the rate matrix Q has at least two different non-zero eigenvalues and two non-zero pairwise distances, the rate matrix Q, pairwise distances and shape parameter α can be identified simultaneously. This conclusion applies to a general time reversible

(GTR) model since it has three non-zero distinct eigenvalues. It also explains why the F81 model is non-identifiable since it only has one non-zero eigenvalue. Moreover, Wu and Susko (2010) provides a theorem stating that the rate matrix, pairwise distance and any arbitrary distribution of the rate factor are identifiable if the rate matrix has at least two distinct non-zero eigenvalues, the expectation of the rate factor is one and the pairwise distribution is available for any distribution. This theorem also serves as a basis for our model when analyzing the identifiability for pairwise sequences when the rate matrix Q has at least two non-zero distinct eigenvalues.

Instead of relying on the number of distinct eigenvalues of the rate matrix Q, Mossel and Roch (2011) have developed a new technique to study the identifiability of large trees. Without the assumption of the gamma distribution of the rate factor, they imply large phylogenies are identifiable with the constraint that the expectation of the rate factor is one. They achieve that by binning sites with similar rates into groups and using a bin with abundant sites to estimate the distance between any two leaves in order to recover the true tree.

In our case, we introduce the rate factor $\gamma = (\gamma_1, \gamma_2, ..., \gamma_k)$ and the latent categorical variable Z to specify which category γ takes. We assume a discrete distribution over γ which is $f = (f_1, f_2, \ldots, f_k)$. The parameters in our model include $\gamma = (\gamma_1, \gamma_2, \ldots, \gamma_k)$ and $f = (f_1, f_2, \ldots, f_k)$. The identifiability of $\theta = (\gamma, f)$ is of particular interest when the rate matrix Q is known as well as the topology of the tree or one pairwise distance.

In this chapter, we have studied the identifiability of $\theta = (\gamma, f)$ for a pair of sequences. We prove that for the F81 model and the JC69 model, $\theta = (\gamma, f)$ is unidentifiable due to the only one non-zero eigenvalues of the rate matrix Q. We apply Wu and Susko (2010)'s conclusions to our situation. We relax the constraint that the expectation of the rate factor is one. We prove that the eigenvalues of the rate matrix are identified if it is unknown but the pairwise distance and the distribution of the rate scalar are unidentifiable for GTR models.

5.3 Identifiability of γ with two categories

For simplicity, we first study the identifiability for only one pair of sequences. We derive the log-likelihood for a pair of sequences with n sites is

$$\begin{aligned} l(\theta) &= \sum_{i=1}^{n} \log \left(\sum_{j=1}^{k} P(Z_i = j)(e^{\gamma_j Q t})_{x_i \to y_i} P(X = x_i) \right) \\ &= \sum_{i=1}^{n} \log \left(\sum_{j=1}^{k} f_j (e^{\gamma_j Q t})_{x_i \to y_i} \pi_{x_i} \right) \qquad (5.1) \end{aligned}$$

with the assumption of independence across different sites. A latent categorical variable Z is introduced to imply which category the rate for a particular site belongs to. The distribution of Z is $f = (f_1, f_2, \ldots, f_k)$. The parameters to be estimated are $\theta = (\gamma, f)$ given the data from pairs of DNA sequences.

To further simplify the problem, consider $k = 2$. Assume $\gamma = (\gamma_1, \gamma_2)$ where $\gamma_1 < \gamma_2$ and $f = (f_1, f_2)$, which is $\theta = (\gamma_1, \gamma_2, f_1, f_2)$ and $l(\theta) = l_0$.

If a model for a pair of sequences is identifiable then there does not exist $\theta^* = (\gamma_1^*, \gamma_2^*, f_1^*, f_2^*)$ (where $\gamma_1^* < \gamma_2^*$) other than θ such that $l(\theta^*) = l_0$.

Assuming $\theta^* = (\gamma_1^*, \gamma_2^*, f_1^*, f_2^*)$ exists to make $l(\theta^*) = l(\theta_0)$, for any

$x_i, y_i \in \{A, G, T, C\}$, then

$$\begin{aligned}&\pi_{x_i} f_1 \exp(\gamma_1 Qt)_{x_i \to y_i} + \pi_{x_i} f_2 exp(\gamma_2 Qt)_{x_i \to y_i}\\ &= \pi_{x_i} f_1^* \exp(\gamma_1^* Qt)_{x_i \to y_i} + \pi_{x_i} f_2^* \exp(\gamma_2^* Qt)_{x_i \to y_i},\end{aligned} \tag{5.2}$$

where π_{x_i} can be cancelled on both sides of the equation. Then it is simplified as

$$\begin{aligned}&f_1 \exp(\gamma_1 Qt)_{x_i \to y_i} + f_2 \exp(\gamma_2 Qt)_{x_i \to y_i}\\ &= f_1^* \exp(\gamma_1^* Qt)_{x_i \to y_i} + f_2^* \exp(\gamma_2^* Qt)_{x_i \to y_i}.\end{aligned} \tag{5.3}$$

Suppose $e^Q = X\text{diag}(e^{d_1}, \ldots, e^{d_p})X^{-1}$, where $d_1, \ldots, d_p$ are the p distinct eigenvalues of Q and $p \leq 4$. It is also known that

$$e^{\gamma Qt} = X\text{diag}(e^{\gamma d_1 t}, \ldots, e^{\gamma d_p t})X^{-1}.$$

In order to make Equation 5.3 to hold, we only need to make sure that

$$M_1 = M_1^*,$$

where

$$M_1 = X \begin{pmatrix} f_1 e^{d_1 \gamma_1 t} + f_2 e^{d_1 \gamma_2 t} & & & \\ & f_1 e^{d_2 \gamma_1 t} + f_2 e^{d_2 \gamma_2 t} & & \\ & & \ddots & \\ & & & f_1 e^{d_p \gamma_1 t} + f_2 e^{d_p \gamma_2 t} \end{pmatrix} X^{-1},$$

$$M_1^* = X \begin{pmatrix} f_1^* e^{d_1 \gamma_1^* t} + f_2^* e^{d_1 \gamma_2^* t} & & & \\ & f_1^* e^{d_2 \gamma_1^* t} + f_2^* e^{d_2 \gamma_2^* t} & & \\ & & \ddots & \\ & & & f_1^* e^{d_p \gamma_1^* t} + f_2^* e^{d_p \gamma_2^* t} \end{pmatrix} X^{-1}.$$

To analyze the identifiability of the model, we are interested to know whether there is $\theta^* = (\gamma_1^*, \gamma_2^*, f_1^*, f_2^*) \neq \theta = (\gamma_1, \gamma_2, f_1, f_2)$ such that the following system of nonlinear equations holds.

$$\begin{cases} f_1 e^{d_1\gamma_1 t} + f_2 e^{d_1\gamma_2 t} = f_1^* e^{d_1\gamma_1^* t} + f_2^* e^{d_1\gamma_2^* t} & (5.4) \\ f_1 e^{d_2\gamma_1 t} + f_2 e^{d_2\gamma_2 t} = f_1^* e^{d_2\gamma_1^* t} + f_2^* e^{d_2\gamma_2^* t} & (5.5) \\ \qquad\qquad \cdots & \\ f_1 e^{d_p\gamma_1 t} + f_2 e^{d_p\gamma_2 t} = f_1^* e^{d_p\gamma_1^* t} + f_2^* e^{d_p\gamma_2^* t} & (5.6) \end{cases}$$

Theorem 3. *Denote $(d_1, d_2, \dots, d_p)$ as p distinct eigenvalues of the rate matrix Q and t as the branch length. For a pair of DNA sequences, if there is a unique solution $\theta = (\gamma, f)$ to the system of Equations meaning that $(\gamma^* = \gamma, f^* = f)$, then the model is identifiable, otherwise, it is unidentifiable.*

In the previous system of equations, there are p equations in total where p is the number of distinct eigenvalues of rate matrix Q satisfying $p \leq 4$. It reflects that the identifiability of the model depends on the number of different eigenvalues of the rate matrix Q.

5.4 Non-identifiability of F81 Model

Proposition 1. *F81 family of rate matrices have one eigenvalue 0 with algebraic multiplicity 1 and the other eigenvalue -1 with algebraic multiplicity 3.*

Proof. By denoting the the diagonal elements as * to make sure the sum of the each row in the Q matrix is zero, where

$$Q \quad = \quad \begin{pmatrix} * & \pi_C & \pi_A & \pi_G \\ \pi_T & * & \pi_A & \pi_G \\ \pi_T & \pi_C & * & \pi_G \\ \pi_T & \pi_C & \pi_A & * \end{pmatrix},$$

the characteristic polynomial of $\chi(\lambda)$ of the 4×4 matrix is

$$\begin{vmatrix} * - \lambda & \pi_C & \pi_A & \pi_G \\ \pi_T & * - \lambda & \pi_A & \pi_G \\ \pi_T & \pi_C & * - \lambda & \pi_G \\ \pi_T & \pi_C & \pi_A & * - \lambda \end{vmatrix}.$$

By doing elementary transformations, $\chi(\lambda)$ can be transformed into

$$\chi(\lambda) = \begin{vmatrix} * - \lambda & \pi_C & 0 & \pi_G \\ \pi_T & * - \lambda & 0 & \pi_G \\ \pi_T & \pi_C & -(1+\lambda) & \pi_G \\ \pi_T & \pi_C & \pi_A + \frac{\pi_A}{\pi_G}(1 + \lambda - \pi_G) & * - \lambda \end{vmatrix}.$$

Similarly, we get

$$\chi(\lambda) = \begin{vmatrix} -(1+\lambda) & \pi_C & 0 & \pi_G \\ \pi_T + \frac{\pi_T}{\pi_C}(1 + \lambda - \pi_C) & * - \lambda & 0 & \pi_G \\ 0 & \pi_C & -(1+\lambda) & \pi_G \\ 0 & \pi_C & \pi_A + \frac{\pi_A}{\pi_G}(1 + \lambda - \pi_G) & * - \lambda \end{vmatrix}.$$

By Laplace expansion along the first column, we get

$$\begin{aligned}
\chi(\lambda) &= -(1+\lambda)\begin{vmatrix} -(1+\lambda-\pi_c) & 0 & \pi_G\left(1+\frac{1+\lambda-\pi_C}{\pi_C}\right) \\ \pi_C & -(1+\lambda) & 0 \\ \pi_C & \pi_A\left(1+\frac{1+\lambda-\pi_G}{\pi_G}\right) & -(1+\lambda) \end{vmatrix} \\
&\quad -\pi_T\left(1+\frac{1+\lambda-\pi_C}{\pi_C}\right)\begin{vmatrix} \pi_C & 0 & 0 \\ \pi_C & -(1+\lambda) & 0 \\ \pi_C & \pi_A\left(1+\frac{1+\lambda-\pi_G}{\pi_G}\right) & -(1+\lambda) \end{vmatrix} \\
&= -(1+\lambda)\left(-(1+\lambda-\pi_C)(1+\lambda)^2+\pi_A(1+\lambda)^2+\pi_G(1+\lambda)^2\right) \\
&\quad -\pi_T\left(1+\lambda\right)^3 \\
&= -(1+\lambda)^3\left(\pi_A+\pi_G+\pi_C-(1+\lambda)\right)-\pi_T\left(1+\lambda\right)^3 \\
&= -(1+\lambda)^3\left(1-\pi_T-(1+\lambda)\right)-\pi_T\left(1+\lambda\right)^3 \\
&= (1+\lambda)^3\left(\lambda+\pi_T\right)-\pi_T\left(1+\lambda\right)^3 \\
&= \lambda\left(1+\lambda\right)^3. \qquad (5.7)
\end{aligned}$$

As a result, $\chi(\lambda)$ has two distinct eigenvalues: $\lambda_1 = 0$ with algebraic multiplicity 1 and $\lambda_2 = -1$ with algebraic multiplicity 3. □

Theorem 4. *Let X and Y be two DNA sequences, Q be any F81 rate matrix of the following form*

$$Q = \begin{pmatrix} * & \pi_C & \pi_A & \pi_G \\ \pi_T & * & \pi_A & \pi_G \\ \pi_T & \pi_C & * & \pi_G \\ \pi_T & \pi_C & \pi_A & * \end{pmatrix}.$$

Denote $\gamma = (\gamma_1, \gamma_2)$ as the rate factor, where $\gamma_1 < \gamma_2$ and $f = (f_1, f_2)$ as the distribution for the latent variable Z specifying which category γ takes. The *branch length t for this pair of sequence is assumed to be 1. Assuming Q is known, the parameter $\theta = (\gamma, f)$ is unidentifiable.*

Proof. Assuming the model is identifiable, $\theta^* = (\gamma_1^*, \gamma_2^*, f_1^*, f_2^*)$ must be equal to $\theta = (\gamma_1, \gamma_2, f_1, f_2)$. For F81 models, we have $d_1 = 0$, $d_2 = -1$.

$$\begin{cases} f_1 e^{0\times\gamma_1 t} + f_2 e^{0\times\gamma_2 t} = f_1^* e^{0\times\gamma_1^* t} + f_2^* e^{0\times\gamma_2^* t} & (5.8) \\ f_1 e^{(-1)\times\gamma_1 t} + f_2 e^{(-1)\gamma_2 t} = f_1^* e^{(-1)\times\gamma_1^* t} + f_2^* e^{(-1)\times\gamma_2^* t} & (5.9) \end{cases}$$

which is

$$\begin{cases} f_1 + f_2 = f_1^* + f_2^* & (5.10) \\ f_1 e^{-\gamma_1 t} + f_2 e^{-\gamma_2 t} = f_1^* e^{-\gamma_1^* t} + f_2^* e^{-\gamma_2^* t} & (5.11) \end{cases}$$

By setting $f_1 = f_1^* = f_2 = f_2^* = 0.5$, $\gamma_1^* = \frac{1}{2}\gamma_1$, t=1, we get $\gamma_2^* = -\log(e^{-\gamma_1} + e^{-\gamma_2} - e^{-0.5\gamma_1})$. This means there is a distinct solution $\theta^* = (\frac{1}{2}\gamma_1, -\log(e^{-\gamma_1} + e^{-\gamma_2} - e^{-0.5\gamma_1}), f_1, f_2)$ which contradicts our assumption that no θ^* exists. We conclude that the model is unidentifiable. □

Example 1. *To illustrate Theorem 4, we provide one rate matrix from F81 family and assume $\pi_A = 0.4$, $\pi_C = 0.3$, $\pi_T = 0.2$ and $\pi_G = 0.1$, $f_1 = 0.5$, $f_2 = 0.5$, $\gamma_1 = 1$, $\gamma_2 = 2$, $t = 1$ and*

$$Q = \begin{pmatrix} * & \pi_C & \pi_A & \pi_G \\ \pi_T & * & \pi_A & \pi_G \\ \pi_T & \pi_C & * & \pi_G \\ \pi_T & \pi_C & \pi_A & * \end{pmatrix} = \begin{pmatrix} -0.8 & 0.3 & 0.4 & 0.1 \\ 0.2 & -0.7 & 0.4 & 0.1 \\ 0.2 & 0.3 & -0.6 & 0.1 \\ 0.2 & 0.3 & 0.4 & -0.9 \end{pmatrix}.$$

The eigenvalues of Q are equal to -1 with algebraic multiplicity 3 and 0 with algebraic multiplicity 1. After setting $f_1^* = 0.5$, $\gamma_1^* = 0.5$, $f_2^* = 0.5$,

$\gamma_2^* = -\log(e^{-1} + e^{-2} - e^{-0.5})$, we get

$$\begin{cases} f_1 + f_2 = f_1^* + f_2^* = 1 & (5.12) \\ 0.5e^{-1} + 0.5e^{-2} = 0.5^*e^{-0.5} + 0.5e^{\log(e^{-1}+e^{-2}-e^{0.5})} & \\ \qquad = 0.5^*e^{-1} + 0.5e^{-2}. & (5.13) \end{cases}$$

Following from Equation 5.4 to Equation 5.6, this model is unidentifiable since $\theta = (1, 2, 0.5, 0.5)$ and $\theta^* = (0.5, -\log(e^{-1} + e^{-2} - e^{-0.5}), 0.5, 0.5)$ have the same likelihood for a pair of sequences.

Corollary 1. *Let X and Y be two DNA sequences, Q be any JC69 rate matrix of the following form,*

$$Q = \begin{pmatrix} * & \frac{\mu}{4} & \frac{\mu}{4} & \frac{\mu}{4} \\ \frac{\mu}{4} & * & \frac{\mu}{4} & \frac{\mu}{4} \\ \frac{\mu}{4} & \frac{\mu}{4} & * & \frac{\mu}{4} \\ \frac{\mu}{4} & \frac{\mu}{4} & \frac{\mu}{4} & * \end{pmatrix}.$$

Assuming $\gamma = (\gamma_1, \gamma_2)$ (where $\gamma_1 < \gamma_2$) is the rate scalar and $f = (f_1, f_2)$ is its distribution, the rate matrix Q and the branch length are known, the parameter $\theta = (\gamma, f)$ is unidentifiable.

Proof. *JC69 rate matrices are special cases of F81 rate matrices by setting $\pi_A = \pi_G = \pi_T = \pi_C = \frac{\mu}{4}$, the result follows from Theorem 4.*

5.5 Identifiability of Models with at least 3 different eigenvalues

In this section, from the F81 model, $\theta = (\gamma, f)$ is unidentifiable since there are only two equations and there are four unknown quantities ($\gamma_1^*, \gamma_2^*, f_1^*$,

f_2^*). As a result, we can find multiple $\theta^* = (\gamma_1^*, \gamma_2^*, f_1^*, f_2^*)$ to satisfy the conditions. However, it is of interest to explore whether the model is identifiable when the rate matrix Q has three distinct eigenvalues or four eigenvalues. When Q has three different eigenvalues, there will be three equations and four unknown quantities $(\gamma_1^*, \gamma_2^*, f_1^*, f_2^*)$ so that the number of equations is still smaller than the number of unknown quantities. When Q has four different eigenvalues, there will be four equations and four unknown quantities $(\gamma_1^*, \gamma_2^*, f_1^*, f_2^*)$ so that the number of equations is equal to the number of unknown quantities. We prove the non-identifiability of the GTR model under certain conditions on the basis of Wu and Susko (2010)'s work.

5.5.1 F84 Model

Recall that in the F1984 model, the rate matrix is given by

$$Q = \begin{pmatrix} * & (1+k/\pi_Y)\pi_C & \pi_A & \beta\pi_G \\ (1+k/\pi_Y)\pi_T & * & \pi_A & \pi_G \\ \pi_T & \pi_C & * & (1+k/\pi_R)\pi_G \\ \pi_T & \pi_C & (1+k/\pi_R)\pi_A & * \end{pmatrix}.$$

The three eigenvalues are $\lambda_1 = 0$, $\lambda_2 = -\mu$, $\lambda_3 = \lambda_4 = -(1+k)\mu$, where μ is $1/(4\pi_T\pi_C(1+k/\pi_Y) + 4\pi_A\pi_G(1+k/\pi_Y) + 4\pi_Y\pi_R)$.

To analyze the identifiability of the F1984 model, we are interested in finding whether there is $\theta^* = (\gamma_1^*, \gamma_2^*, f_1^*, f_2^*) \neq \theta = (\gamma_1, \gamma_2, f_1, f_2)$ such that the following nonlinear equations hold. If θ^* exists, then the model is unidentifiable.

$$
\begin{cases}
f_1 + f_2 = f_1^* + f_2^* & (5.14) \\
f_1 e^{-\mu\gamma_1 t} + f_2 e^{-\mu\gamma_2 t} = f_1^* e^{-\mu\gamma_1^* t} + f_2^* e^{-\mu\gamma_2^* t} & (5.15) \\
f_1^* e^{-(1+k)\mu\gamma_1^* t} + f_2^* e^{-(1+k)\mu\gamma_2^* t} = f_1^* e^{-(1+k)\mu\gamma_1^* t} + f_2^* e^{-(1+k)\mu\gamma_2^* t} & (5.16)
\end{cases}
$$

5.5.2 HKY85 Model

Hasegawa et al. (1985) prove that in HKY85 model, Q has four distinct eigenvalues. If Q is as below, the ratio of α and β is the ratio of transition and tranversion. There are four distinct eigenvalues of the rate matrix

$$
\begin{pmatrix}
* & \alpha\pi_C & \beta\pi_A & \beta\pi_G \\
\alpha\pi_T & * & \beta\pi_A & \beta\pi_G \\
\beta\pi_T & \beta\pi_C & * & \alpha\pi_G \\
\beta\pi_T & \beta\pi_C & \alpha\pi_A & *
\end{pmatrix},
$$

where $\lambda_1 = 0$, $\lambda_2 = -\beta$, $\lambda_3 = -(\pi_Y\beta + \pi_R\alpha)$, $\lambda_4 = (\pi_Y\alpha + \pi_R\beta)$.

The identifiability of HKY85 model lies in whether we are able to find $\theta^* = (\gamma_1^*, \gamma_2^*, f_1^*, f_2^*) \neq \theta = (\gamma_1, \gamma_2, f_1, f_2)$ such that the following nonlinear equations hold.

$$
\begin{cases}
f_1 + f_2 = f_1^* + f_2^* & (5.17) \\
f_1 e^{-\lambda_2\gamma_2 t} + f_2 e^{-\lambda_2\gamma_2 t} = f_1^* e^{-\lambda_2\gamma_1^* t} + f_2^* e^{-\lambda_2\gamma_2^* t} & (5.18) \\
f_1^* e^{-\lambda_3\gamma_1^* t} + f_2^* e^{-\lambda_3\gamma_2^* t} = f_1^* e^{-\lambda_3\gamma_1^* t} + f_2^* e^{-\lambda_3\gamma_2^* t} & (5.19) \\
f_1 e^{-\lambda_4\gamma_1 t} + f_2 e^{-\lambda_4\gamma_2 t} = f_1^* e^{-\lambda_4\gamma_1^* t} + f_2^* e^{-\lambda_4\gamma_2^* t} & (5.20)
\end{cases}
$$

These conditions also apply to GTR models since they also have four distinct eigenvalues as HKY85 model.

We cite Wu and Susko (2010)'s proved Theorem 5 as below to address the identifiability issue for our model.

Theorem 5. *Consider the GTR model with unknown rate matrix Q, which has an unknown non-zero stationary frequencies. The rate factor γ is described by an arbitrary distribution ω, $E_\omega(\gamma) = 1$. If pairwise distributions are available for any distance, then the rate distribution, the pairwise distance d and the rate matrix Q are identifiable.*

The details of proof for Theorem 5 are provided by Wu and Susko (2010). In our case, the rate factor γ comes from an arbitrary discrete distribution without the constraint of $E_\omega(\gamma) = 1$ so we generalize Theorem 5 into Theorem 6.

Theorem 6. *For any GTR model with unknown rate matrix Q, which has at least two distinct eigenvalues and unknown non-zero stationary frequencies. The rate factor γ is described by an arbitrary distribution ω. If pairwise distributions are available for any distance, the eigenvalues of the rate matrix Q are identifiable but the rate distribution and the pairwise distance are unidentifiable. However, if two choices of distance, rate distribution (d, ω) and $(\tilde{d}, \tilde{\omega})$ can lead to the same pairwise distribution, the following equation holds*

$$d \times E_\omega(\gamma) = \tilde{d} \times E_{\tilde{\omega}}(\tilde{\gamma}).$$

Proof. The proof can be obtained using the same technique when Wu and Susko (2010) proved their "Theorem 2" in the original paper. For reader's convenience, we used the same notations as the original paper. Assuming the rate factor γ has a distribution ω, then the moment generating function is

$$M(t) = E_\omega(e^{t\gamma})$$

for any $t \leq 0$. Assuming (d, ω, Q) and $(\tilde{d}, \tilde{\omega}, \tilde{Q})$ are two different choices of the pairwise distance, distribution of the rate factor and rate matrix. As denoted before, M and $\tilde{M}$ are the two moment generating functions corresponding to γ and $\tilde{\gamma}$. The rate matrix Q has n distinct eigenvalues λ_i where $i \in \{1, 2, \ldots, n\}$. Moreover, Q can be represented as $X\text{diag}(\lambda_1, \lambda_2, \ldots, \lambda_n)X^{-1}$ and $\tilde{Q}$ can be represented as $\tilde{X}\text{diag}(\tilde{\lambda_1}, \tilde{\lambda_2}, \ldots, \tilde{\lambda_n})\tilde{X}^{-1}$. Let c_1 denote $E_\omega(\gamma)$, c_2 denote $E_{\tilde{\omega}}(\tilde{\gamma})$ and c denote the ratio of c_1 and c_2. From the properties of moment generating functions, we have

$$\frac{M'(0)}{\tilde{M}'(0)} = \frac{E_\omega(\gamma)}{E_{\tilde{\omega}}(\tilde{\gamma})} = \frac{c_1}{c_2} = c.$$

It follows from the proof of Theorem 2 in the original paper of Wu and Susko (2010) that for any $v \leq 0$, we have $M(v) = \tilde{M}(cv)$. When the pairwise distribution is the same for two choices of (d, μ, Q) and $(\tilde{d}, \tilde{\mu}, \tilde{Q})$, then

$$\tilde{d} = \tilde{\lambda_i}^{-1}\tilde{M}^{-1}(M(\lambda_i d)).$$

Since $M(v) = \tilde{M}(cv)$, then

$$\tilde{d} = \tilde{\lambda_i}^{-1}\tilde{M}^{-1}(M(\lambda_i d)) = c\frac{\lambda_i}{\tilde{\lambda_i}}d,$$

$$\frac{\tilde{\lambda_i}}{\lambda_i} = c \times \frac{d}{\tilde{d}} = c \times k.$$

From the proof of Theorem 1(a) by Wu and Susko (2010), since the unknown stationary distribution Π can be obtained using the 1-taxa marginalization and the rate matrices are scaled so that trace(ΠQ) $= -1$, then,

$$\begin{aligned} -1 &= \text{trace}\left(\Pi\tilde{X}\text{diag}(\tilde{\lambda_1}, \tilde{\lambda_2}, \ldots, \tilde{\lambda_p})\tilde{X}^{-1}\right) \\ &= \text{trace}\left(\Pi X\text{diag}(\lambda_1, \lambda_2, \ldots, \lambda_p)X^{-1}\right) = -ck, \end{aligned}$$

so that

$$ck = \frac{E_\omega(\gamma)}{E_{\tilde{\omega}}(\gamma)} \times \frac{d}{\tilde{d}} = 1,$$

where $\lambda_i = \tilde{\lambda_i}$ for any $i \in \{1, 2, \ldots, n\}$. □

5.6 Simulation Results

To study the identifiability of γ and its distribution f, we select HKY85 model since it has three distinct nonzero eigenvalues. We are interested to see if the rate matrix is known, whether it is possible to recover the distribution of parameters γ using a pair of sequences.

Using the HKY85 model, the rate matrix is

$$Q = \mu * \begin{pmatrix} * & \pi_C & \kappa\pi_G & \pi_T \\ \pi_A & * & \pi_G & \kappa\pi_T \\ \kappa\pi_A & \pi_C & * & \pi_T \\ \pi_A & \kappa\pi_C & \pi_G & * \end{pmatrix}.$$

The normalization constant is $\mu = 1/\left\{\sum_{i\{A,C,G,T\}} \pi_i Q_{ii}\right\}$. After setting $\kappa = 5$ and the stationary distribution as $\pi = (\pi_A, \pi_C, \pi_G, \pi_T) = (0.4, 0.3, 0.2, 0.1)$, we obtain the rate matrix as below

$$Q_0 = \begin{array}{c} \\ A \\ C \\ G \\ T \end{array} \begin{array}{c} \begin{array}{cccc} A & C & G & T \end{array} \\ \begin{pmatrix} -0.89 & 0.19 & 0.63 & 0.06 \\ 0.25 & -0.70 & 0.13 & 0.32 \\ 1.27 & 0.19 & -1.52 & 0.06 \\ 0.25 & 0.95 & 0.13 & -1.33 \end{pmatrix} \end{array}.$$

We generate a pair of sequences with 50000 sites from the previous HKY85 model with Q_0 as the rate matrix and the branch length is 0.2. The rate factor is $\gamma = (0, 1)$ and its distribution is $f = (0.2, 0.8)$. This means that among 50000 sites, approximately 20 percent of the sites stay unchanged and 80 percent of the sites change according to the rate matrix Q_0. Using the EM algorithm to learn γ and f for a pair of sequences, assuming we know Q_0, we are interested to see whether we can recover the reference held-out parameters $\gamma = (0, 1)$ and its distribution $f = (0.2, 0.8)$ using the generated dataset. If the estimate of γ and f is close to γ and f, it is promising that γ and its distribution f are identifiable provided that the rate matrix is known. If the estimate of $\hat{\gamma}$ and $\hat{f}$ are very different from γ and f, but these two choices of parameters can produce the same or very similar likelihood, this will provide some empirical support for the non-identifiability for HKY85 model. We replicate simulating sequences with $\gamma = (0, 1)$ and $f = (0.2, 0.8)$, then make inference about γ and f with different initial values of γ and f denoted as $\gamma_{initial}$ and $f_{initial}$. For 500 times, the average of $\hat{\gamma}$ and $\hat{f}$ are summarized as below.

Table 5.1: Estimating γ and f with $\gamma_{initial}$ and $f_{initial}$

$\gamma_{initial}$	$f_{initial}$	$\hat{\gamma}$	$\hat{f}$	$SD(\hat{\gamma})$	$SD(\hat{f})$
(0, 1)	(0.2, 0.8)	(0, 0.96)	(0.178, 0.822)	(0, 0.069)	(0.048, 0.048)
(0, 0.5)	(0.5, 0.5)	(0, 1.01)	(0.208, 0.792)	(0, 0.093)	(0.059, 0.059)
(0.4, 0.5)	(0.5, 0.5)	(0.546, 1.011)	(0.497, 0.503)	(0.176, 0.201)	(0.002, 0.002)
(0.2, 0.8)	(0.5, 0.5)	(0.510, 1.04)	(0.485, 0.515)	(0.179, 0.197)	(0.007, 0.007)
(0.2, 0.8)	(0.3, 0.7)	(0.417, 0.935)	(0.293, 0.707)	(0.260, 0.136)	(0.008, 0.008)

When the initial values $\gamma_{initial}$ and $f_{initial}$ are equal to the reference held-out parameters γ and f, the estimated parameters $\hat{\gamma}$ and $\hat{f}$ are close to the parameters we use to simulate the data. The standard error of $\hat{\gamma}$ and $\hat{f}$ represents the variation of $\hat{\gamma}$ and $\hat{f}$ out of 500 replications. When only one category of $\gamma_{initial}$ is the same as γ which is zero, $\hat{\gamma}$ and $\hat{f}$ are still close to γ and f. When $\gamma_{initial}$ and $f_{initial}$ are very different from γ and f, $\hat{\gamma}$ and $\hat{f}$ are no longer close to γ and f. Moreover, $\hat{f}$ highly depends on $f_{initial}$ which is reflected from the last three rows of Table 5.1.

To investigate the identifiability of the problem, we are interested to check whether the negative log-likelihood given $\hat{\gamma}$ and $\hat{f}$ are the same as the negative log-likelihood given γ and f.

Table 5.2: Average of nllk of the generated dataset given different estimates

$\hat{\gamma}$	$\hat{f}$	$n\bar{l}lk_{<\hat{\gamma},\hat{f}>}$	$n\bar{l}lk_{<\gamma,f>}$	Range	Times
(0.417, 0.935)	(0.29, 0.71)	88839	88841	(-7.30, 0.03)	494
(0.510, 1.04)	(0.49, 0.51)	88855	88856	(-8.70, 0.06)	492
(0.546, 1.011)	(0.50, 0.50)	88850	88852	(-6.99, 0.05)	491
(0, 1.011)	(0.21, 0.79)	88825	88826	(-6.75, 1.70)	421
(0, 0.96)	(0.18, 0.82)	88839	88834	(-6.28, 2.42)	411

In Table 5.2, “Average of nllk” represents the average of negative log-likelihood out of 500 replications. “Times” denotes the number of times that the negative log-likelihood of the generated dataset given $\hat{\gamma}$ and $\hat{f}$ is smaller than the one given γ and f used to generate the dataset. “Range” denotes the range of the difference between the negative log-likelihood given $(\hat{\gamma}, \hat{f})$ and (γ, f).

Based on these simulations, we conjecture that γ and f are not identifiable in this setup. We also produce two pictures to help visualize that. We generate a pair of sequences with $\gamma = (0.6, 1)$ and $f = (0.2, 0.8)$.

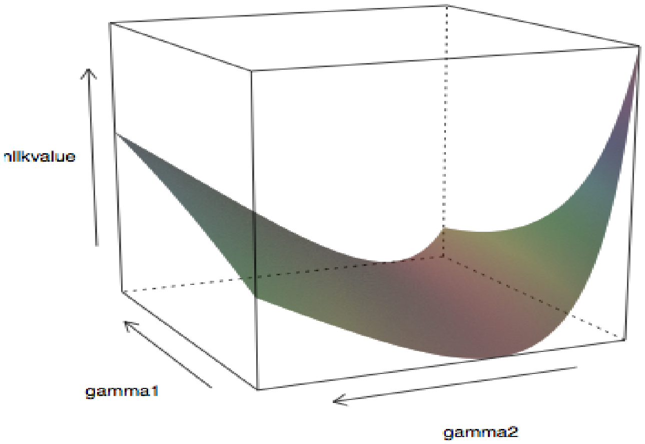

Figure 5.1: negative log-likelihood given γ by grid search

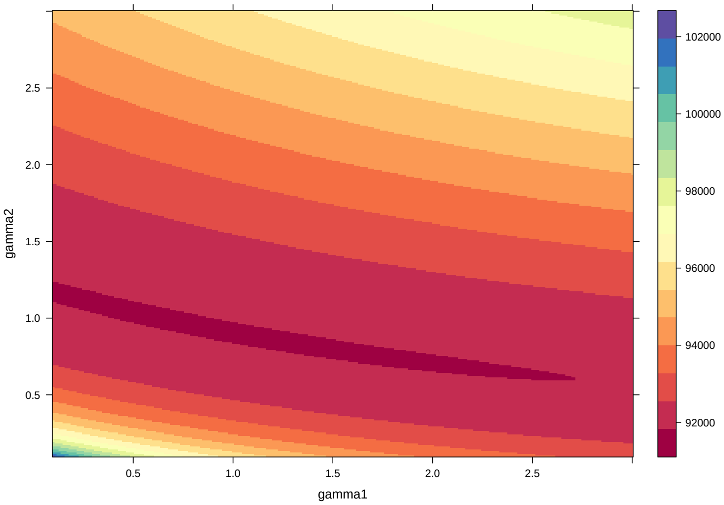

Figure 5.2: level plot of negative log-likelihood given γ by grid search

Figure 5.1 shows the negative log-likelihood given different rate factors. In this figure, “gamma1” denotes the first category of γ and “gamma2” denotes the second category of γ. It shows on the bottom of the plot, the plane is flat which is coloured by “red”. To check this, we produce Figure 5.2 by projecting the 3-d plot to a 2-d plane. From the curve in the plot, it reflects that different combinations of the two categories of γ could produce similar negative log-likelihood. This also serves as the empirical support that the HKY85 model for a pair of sequences might not be identifiable.

Chapter 6

Data Analysis

6.1 Data Description

The dataset comes from the original paper by Shang, Xu, Ozer, and Gutell (2012). There are 1028 ribosomal RNA sequences. Each sequence has 4907 sites of 16S gene and is sampled from eucarya mitochondria. The topology and branch lengths of the tree are first estimated by maximum likelihood methods.

Let $X = \{x_{ij}\}$ be the aligned nucleotide sequences, where $i \in \{1, 2, \ldots, 1028\}$, $j \in \{1, 2, ..., 4907\}$. Each column of the data matrix $x_j = \{x_{1,j}, ..., x_{1028,j}\}$ specifies the nucleotides for the 1028 sequences at the jth site. Each row of the data matrix $x_i = \{x_{i,1}, ..., x_{i,4907}\}$ represents all the nucleotides of the ith DNA sequence.

$$X = \begin{pmatrix} x_{1,1} & x_{1,2} & \cdots & x_{1,4907} \\ x_{2,1} & x_{2,2} & \cdots & x_{2,4907} \\ \vdots & \vdots & \ddots & \vdots \\ x_{1028,1} & x_{1028,2} & \cdots & x_{1028,4907} \end{pmatrix}$$

However, a considerable fraction of x_{ij} are considered missing. In phylogenetics, missing data are not usually due to missing measurements, but are rather artificially introduced as a preprocessing step called sequence alignment. In order to match the sequences with the same length, part of the missing data denoted as “-” in some sequences are “deletions or “insertions” in which nucleotides are deleted or inserted. Below we show a histogram of the percentage of the sites with different numbers of missing data.

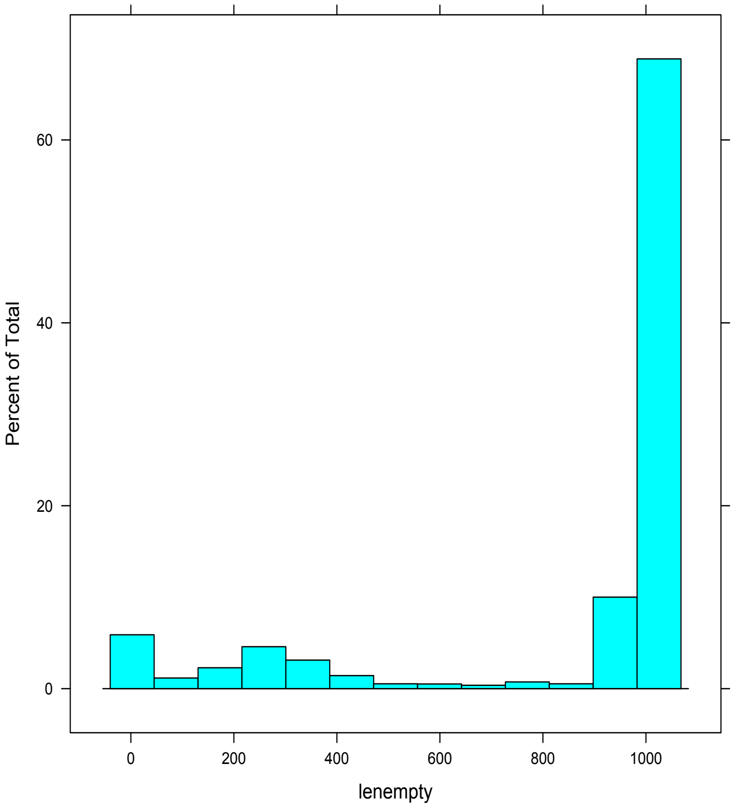

Figure 6.1: percentage of sites with different numbers of missing data

For sixty percent of the sites, the states of more than 1000 sequences on each of this site is missing. For example, when we check the data, on the 4907th site, only the state of one sequence is known and the states of all other sequences at this site are missing. This kind of sites are not used when analyzing the data.

6.2 Exploratory Data Analysis

Here we explore whether the rates at different sites are the same or not. We use the sites with the least missing data in the original dataset. In total, there are 44 sites at which only the states of two sequences out of 1028 sequences are missing.

In this chapter, we use the K80 model. It is assumed that the stationary frequencies of the nucleotides π_x are uniform and the ratio of transition over transversion κ is equal to 1.5. Then the rate matrix Q used in this dataset is

$$Q = \begin{pmatrix} * & \pi_C & \kappa\pi_G & \pi_T \\ \pi_A & * & \pi_G & \kappa\pi_T \\ \kappa\pi_A & \pi_C & * & \pi_T \\ \pi_A & \kappa\pi_C & \pi_G & * \end{pmatrix} = \begin{pmatrix} -0.875 & 0.25 & 0.375 & 0.25 \\ 0.25 & -0.875 & 0.25 & 0.375 \\ 0.375 & 0.25 & -0.875 & 0.25 \\ 0.25 & 0.375 & 0.25 & -0.875 \end{pmatrix} \times \mu,$$

where $\mu = 1.142857$.

We would like to explore whether the rates at these sites are the same or not. We calculate the likelihood for each of these sites under different rate factors with one category ranging from 0.01 to 1 with step size 0.01. We denote the rate factor as γ and show three figures of the log-likelihood with respect to the rate factor at three different sites.

The patterns of the three figures are similar since they are all unimodal and the log-likelihood increases as γ increases first and decreases after a certain point. We provide a histogram of the estimate of γ by grid search for the selected sites. The histogram reflects that the distribution of $\hat{\gamma}$ at

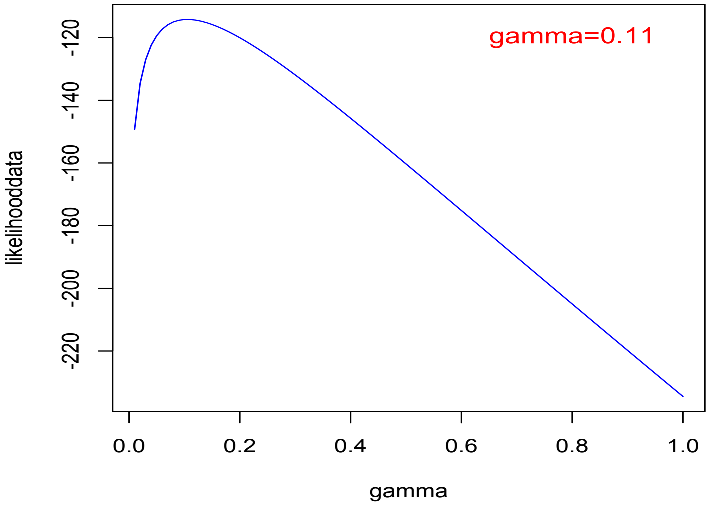

Figure 6.2: log-likelihood w.r.t γ at site “1”

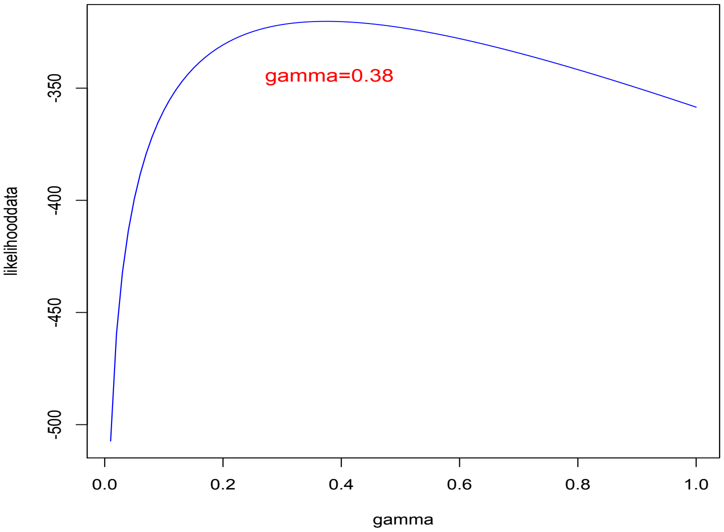

Figure 6.3: log-likelihood w.r.t γ at site “2”

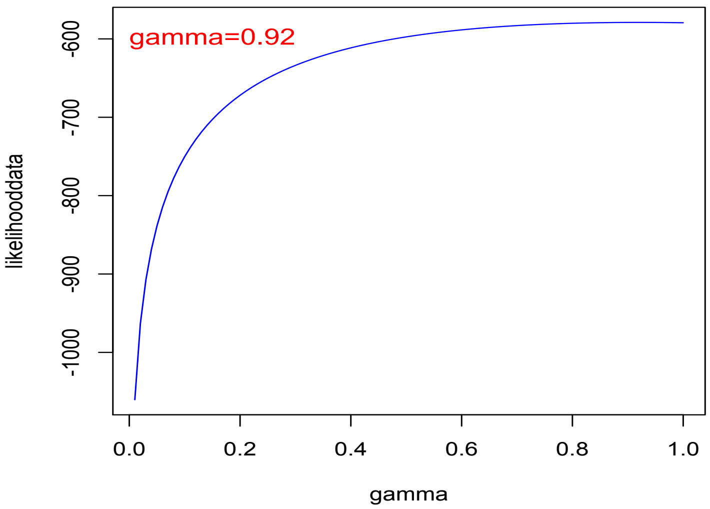

Figure 6.4: log-likelihood w.r.t γ at site "3"

which the maximum likelihood is achieved by grid search method. Note that more than sixty percent of the sites evolve slowly relative to the other ones.

6.3 Data Analysis via an EM Algorithm

The fixed quantities include the topology of the tree and the branch lengths. The parameters to be estimated are the rate factor γ and its distribution f. Without introducing the rate factor, the likelihood of the tree is -381560.755. Using an EM algorithm with a penalty term which is the sum of γ^2, we provide the results as below. As to the details of the penalty terms, please refer to Chapter 4. We get the estimates by removing the sites where the states of 1027 sequences are missing for a single site since those

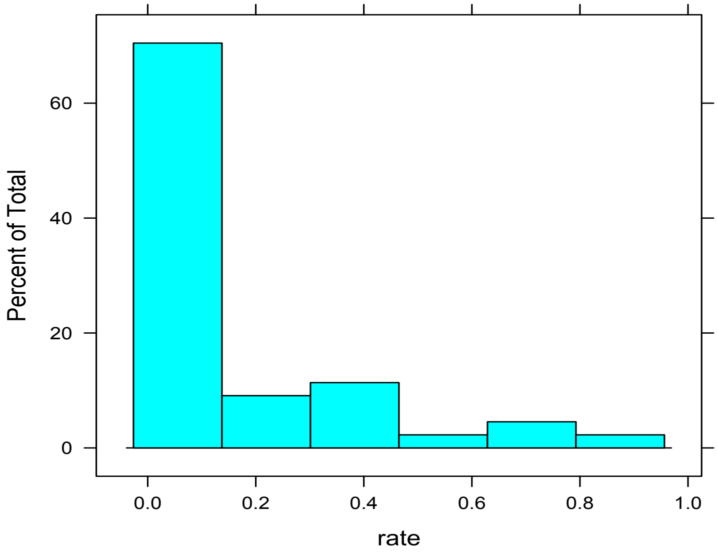

Figure 6.5: Histogram of $\hat{\gamma}_{MLE}$ by grid search

sites only have the state of one sequence so that it does not provide much information.

Table 6.1: Estimating γ and f with different $\gamma_{initial}$ and $f_{initial}$ with part of alignments

$\gamma_{initial}$	$f_{initial}$	$\hat{\gamma}$	$\hat{f}$	$nllk_{\{\hat{\gamma},\hat{f}\}}$
0.5	1	0.381	1	321374
1	1	0.381	1	321374
(0.4, 1.4)	(0.5, 0.5)	(0.236, 3.531)	(0.488, 0.512)	303971
(0.16, 0.6, 1.2)	$(\frac{1}{3}, \frac{1}{3}, \frac{1}{3})$	(0.116, 0.926, 2.602)	(0.238, 0.420, 0.342)	290146
(0.12, 0.15, 0.6, 0.9)	$(\frac{1}{4}, \frac{1}{4}, \frac{1}{4}, \frac{1}{4})$	(0.038, 0.160, 0.591, 3.148)	(0.087, 0.125, 0.380, 0.408)	287676
(0.12, 0.15, 0.6, 0.9, 1.2)	$(\frac{1}{5}, \frac{1}{5}, \frac{1}{5}, \frac{1}{5}, \frac{1}{5})$	(0.389, 0.152, 0.393,0.961, 2.727)	(0.087, 0.104, 0.166, 0.271, 0.372)	285772

In order to perform the model selection, we calculate the the Akaike information criterion (AIC) and Bayesian information criterion (BIC) for the five models.

Table 6.2: AIC and BIC for different Models

#of categories	AIC	BIC
1	642756	642782
2	607948	607988
3	580302	580368
4	575366	575413
5	571562	571681

Since both the AIC and BIC keep increasing for the five models as the number of parameters increases, we suspect it under-penalize model complexity. The reason is that the dependence between sequences violates the assumption of AIC and BIC so that a new model selection method is needed. To find a new model selection method is part of our future work.

Chapter 7

Conclusions and Future Work

In this thesis, we propose a new computationally attractive model. Using an EM algorithm, we can model the rates over sites on DNA sequences with a discrete distribution for thousands of sequences. We also analyze the identifiability for the rate factor coming from a discrete distribution for the first time. We prove a general condition of the identifiability of our model. Based on that, we prove the non-identifiability of F81 model. We also prove the non-identifiability of GTR model under certain conditions.

For future work, we are interested to explore new model selection methods to evaluate our models. Moreover, we would like to propose new methods to estimate the rate matrix Q and the rate factor γ and its distribution simultaneously. Finally, we are interested to analyze the identifiability of large phylogenies.

Bibliography

Elizabeth S Allman, Cécile Ané, and John A Rhodes. Identifiability of a Markovian model of molecular evolution with gamma-distributed rates. *Advances in Applied Probability*, pages 229–249, 2008.

David Bryant, Nicolas Galtier, and Marie-Anne Poursat. Likelihood calculation in molecular phylogenetics. *Mathematics of Evolution and Phylogeny, cáp*, 2:33–62, 2005.

Joseph Felsenstein. Evolutionary trees from DNA sequences: a maximum likelihood approach. *Journal of molecular evolution*, 17(6):368–376, 1981.

Joseph Felsenstein. *Inferring phylogenies*, volume 2. Sinauer Associates Sunderland, 2004.

Dan Graur and Wen-Hsiung Li. *Fundamentals of molecular evolution*, volume 2. Sinauer Associates Sunderland, 2000.

Masami Hasegawa, Hirohisa Kishino, and Taka-aki Yano. Dating of the human-ape splitting by a molecular clock of mitochondrial DNA. *Journal of molecular evolution*, 22(2):160–174, 1985.

Thomas H Jukes and Charles R Cantor. Evolution of protein molecules. 1969.

Motoo Kimura. A simple method for estimating evolutionary rates of base substitutions through comparative studies of nucleotide sequences. *Journal of molecular evolution*, 16(2):111–120, 1980.

Elchanan Mossel and Sebastien Roch. Identifiability and inference of nonparametric rates-across-sites models on large-scale phylogenies. *Journal of Mathematical Biology*, pages 1–31, 2011.

Lei Shang, Weijia Xu, Stuart Ozer, and Robin R Gutell. Structural constraints identified with covariation analysis in ribosomal RNA. *PloS one*, 7(6):e39383, 2012.

Mike Steel. A basic limitation on inferring phylogenies by pairwise sequence comparisons. *Journal of theoretical biology*, 256(3):467–472, 2009.

RM Wilcox. Exponential operators and parameter differentiation in quantum physics. *Journal of Mathematical Physics*, 8:962, 1967.

Jihua Wu and Edward Susko. Rate-variation need not defeat phylogenetic inference through pairwise sequence comparisons. *Journal of theoretical biology*, 263(4): 587–589, 2010.

Ziheng Yang. Maximum-likelihood estimation of phylogeny from DNA sequences when substitution rates differ over sites. *Molecular Biology and Evolution*, 10(6): 1396–1401, 1993.

Ziheng Yang. Maximum likelihood phylogenetic estimation from DNA sequences with variable rates over sites: approximate methods. *Journal of Molecular evolution*, 39(3):306–314, 1994.

Ziheng Yang. Among-site rate variation and its impact on phylogenetic analyses. *Trends in Ecology & Evolution*, 11(9):367–372, 1996.

Appendix A

First Appendix

Theorem 1.

$$\lim_{t_n \to t} \frac{(e^{t_n Q})_{i,j} - (e^{tQ})_{i,j}}{t_n - t} = (Qe^{tQ})_{i,j}$$

Proof. If $A = [a_{ij}]$ is a $p \times p$ matrix, the matrix exponential of A is $e^A = \sum_{j=0}^{\infty} A^j/j!$. Assume A has p distinct eigenvalues $d_1, d_2, \ldots, d_p$ so that X is the $p \times p$ matrix where the jth column is a right eigenvector of unit length corresponding to d_j. Since $A = XDX^{-1}$,

$$e^{tA} = X\text{diag}(e^{d_1 t}, \ldots, e^{d_p t})X^{-1}$$

$$e^{tA} = \begin{pmatrix} x_{11}e^{d_1 t} & x_{12}e^{d_2 t} & \ldots & x_{1p}e^{d_p t} \\ x_{21}e^{d_1 t} & x_{22}e^{d_2 t} & \ldots & x_{2p}e^{d_p t} \\ \vdots & \vdots & \ddots & \vdots \\ x_{p1}e^{d_1 t} & x_{p2}e^{d_2 t} & \ldots & x_{pp}e^{d_p t} \end{pmatrix} \begin{pmatrix} x^*_{11} & x^*_{12} & \ldots & x^*_{1p} \\ x^*_{21} & x^*_{22} & \ldots & x^*_{2p} \\ \vdots & \vdots & \ddots & \vdots \\ x^*_{p1} & x^*_{p2} & \ldots & x^*_{pp} \end{pmatrix}$$

where

$$X^{-1} = \begin{pmatrix} x^*_{11} & x^*_{12} & \ldots & x^*_{1p} \\ x^*_{21} & x^*_{22} & \ldots & x^*_{2p} \\ \vdots & \vdots & \ddots & \vdots \\ x^*_{p1} & x^*_{p2} & \ldots & x^*_{pp} \end{pmatrix} \quad X = \begin{pmatrix} x_{11} & x_{12} & \ldots & x_{1p} \\ x_{21} & x_{22} & \ldots & x_{2p} \\ \vdots & \vdots & \ddots & \vdots \\ x_{p1} & x_{p2} & \ldots & x_{pp} \end{pmatrix}$$

$$(e^{tA})_{i,j} = \sum_{k=1}^{p} x_{ik}e^{d_k t}x^*_{ki} = \sum_{k=1}^{p} x_{ik}x^*_{ki}e^{d_k t}$$

$$\lim_{t_n \to t} \frac{(e^{t_n Q})_{i,j} - (e^{tQ})_{i,j}}{t_n - t} = \frac{\partial (e^{tA})_{i,j}}{\partial t} = \frac{\partial (\sum_{k=1}^{p} x_{ik}x^*_{ki}e^{d_k t})}{\partial t} = \sum_{k=1}^{p} x_{ik}x^*_{ki}d_k e^{d_k t}$$

$$\begin{aligned}
Ae^{tA} &= X\mathrm{diag}(d_1, d_2, \ldots, d_p)X^{-1}X\mathrm{diag}(e^{d_1 t}, \ldots, e^{d_p}t)X^{-1} \\
&= X\mathrm{diag}(d_1, \ldots, d_p)\mathrm{diag}(e^{d_1 t}, \ldots, e^{d_p t})X^{-1} \\
&= \begin{pmatrix} x_{11}d_1 & x_{12}d_2 & \ldots & x_{1p}d_p \\ x_{21}d_1 & x_{22}d_2 & \ldots & x_{2p}d_p \\ \vdots & \vdots & \ddots & \vdots \\ x_{p1}d_1 & x_{p2}d_2 & \ldots & x_{pp}d_p \end{pmatrix} \begin{pmatrix} e^{d_1 t} & 0 & \ldots & 0 \\ 0 & e^{d_2 t} & \ldots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \ldots & e^{d_n t} \end{pmatrix} \begin{pmatrix} x^*_{11} & x^*_{12} & \ldots & x^*_{1p} \\ x^*_{21} & x^*_{22} & \ldots & x^*_{2p} \\ \vdots & \vdots & \ddots & \vdots \\ x^*_{p1} & x^*_{p2} & \ldots & x^*_{pp} \end{pmatrix} \\
&= \begin{pmatrix} x_{11}d_1e^{d_1 t} & x_{12}d_2e^{d_2 t} & \ldots & x_{1p}d_pe^{d_p t} \\ x_{21}d_1e^{d_1 t} & x_{22}d_2e^{d_2 t} & \ldots & x_{2p}d_pe^{d_p t} \\ \vdots & \vdots & \ddots & \vdots \\ x_{p1}d_1e^{d_1 t} & x_{p2}d_2e^{d_2 t} & \ldots & x_{pp}d_pe^{d_p t} \end{pmatrix} \begin{pmatrix} x^*_{11} & x^*_{12} & \ldots & x^*_{1p} \\ x^*_{21} & x^*_{22} & \ldots & x^*_{2p} \\ \vdots & \vdots & \ddots & \vdots \\ x^*_{p1} & x^*_{p2} & \ldots & x^*_{pp} \end{pmatrix}
\end{aligned}$$

Then

$$(Ae^{tA})_{ij} = \sum_{k=1}^{p} x_{ik}x^*_{ki}d_ke^{d_k t} = \frac{\partial(\sum_{k=1}^{p} x_{ik}x^*_{ki}e^{d_k t})}{\partial t} = \lim_{t_n \to t} \frac{(e^{t_n Q})_{i,j} - (e^{tQ})_{i,j}}{t_n - t}$$

$\square$

Appendix B

Second Appendix

Theorem 1.

$$
\begin{aligned}
&\frac{\partial}{\partial\gamma}(\exp(\gamma Q_1)_{m\to n}\exp(\gamma Q_2)_{a\to b})\\
&=[\frac{\partial}{\partial\gamma}(\exp(\gamma Q_1))]_{m\to n}\exp(\gamma Q_2)_{a\to b}+\exp(\gamma Q_1)_{m\to n}[\frac{\partial}{\partial\gamma}(\exp(\gamma Q_2))]_{a\to b}\\
&=(Q_1\exp(\gamma Q_1))_{m\to n}\exp(\gamma Q_2)_{a\to b}+\exp(\gamma Q_1)(Q_2\exp(\gamma Q_2))_{a\to b}
\end{aligned}
$$

The proof of this theorem is provided in the Appendix.

Proof. It has been proved that,

$$\lim_{t_n\to t}\frac{(e^{t_nQ})_{i,j}-(e^{tQ})_{i,j}}{t_n-t}=(Qe^{tQ})_{i,j}$$

When we get derivative wrt t for the element in the ith row and jth column of the matrix $(e^{tA})_{ij}$ is equivalent to getting derivative wrt t for the matrix first and then take the element in the ith row and jth column of the matrix. Then

$$
\begin{aligned}
&\frac{\partial}{\partial\gamma}(\exp(\gamma Q_1)_{m\to n}\exp(\gamma Q_2)_{a\to b})\\
&=(\frac{\partial}{\partial\gamma}\exp(\gamma Q_1)_{m\to n})\exp(\gamma Q_2)_{a\to b}+\exp(\gamma Q_1)_{m\to n}(\frac{\partial}{\partial\gamma}\exp(\gamma Q_2)_{a\to b})\\
&=(\frac{\partial}{\partial\gamma}\exp(\gamma Q_1))_{m\to n}\exp(\gamma Q_2)_{a\to b}+\exp(\gamma Q_1)_{m\to n}(\frac{\partial}{\partial\gamma}\exp(\gamma Q_2))_{a\to b} \qquad \text{(B.1)}
\end{aligned}
$$

□