THE EXPLORATION OF EFFECT OF MODEL MISSPECIFICATION AND DEVELOPMENT OF AN ADEQUACY-TEST FOR SUBSTITUTION MODEL IN PHYLOGENETICS

by

Wei Chen

Submitted in partial fulfillment of the requirements
for the degree of Master of Science

at

Dalhousie University
Halifax, Nova Scotia
November 2012

DALHOUSIE UNIVERSITY

DEPARTMENT OF MATHEMATICS AND STATISTICS

The undersigned hereby certify that they have read and recommend to the Faculty of Graduate Studies for acceptance a thesis entitled "THE EXPLORATION OF EFFECT OF MODEL MISSPECIFICATION AND DEVELOPMENT OF AN ADEQUACY-TEST FOR SUBSTITUTION MODEL IN PHYLOGENETICS" by Wei Chen in partial fulfillment of the requirements for the degree of Master of Science.

Dated: November 6, 2012

Co-supervisors: _____

_____

Readers: _____

_____

DALHOUSIE UNIVERSITY

DATE: November 6, 2012

AUTHOR: Wei Chen

TITLE: THE EXPLORATION OF EFFECT OF MODEL MISSPECIFICATION AND DEVELOPMENT OF AN ADEQUACY-TEST FOR SUBSTITUTION MODEL IN PHYLOGENETICS

DEPARTMENT OR SCHOOL: Department of Mathematics and Statistics

DEGREE: M.Sc. CONVOCATION: May YEAR: 2013

Signature of Author

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

ABSTRACT

It is possible that the maximum likelihood method can give an inconsistent result when the DNA sequences are generated under a tree topology which is in the Felsentein Zone and analyzed with a misspecified model. Therefore, it is important to select a good substitution model. This thesis first explores the effects of different degrees and types of model misspecification on the maximum likelihood estimates. The results are presented for tree selection and branch length estimates based on simulated data sets. Next, two Pearson's goodness-of-fit tests are developed based on binning of site patterns. These two tests are used for testing the adequacy of substitution models and their performances are studied on both simulated data sets and empirical data.

LIST OF ABBREVIATIONS AND SYMBOLS USED

Roman symbol	Description
a	short branch 1
b	long branch 1
c	short branch 3
d	long branch 2
e	internal branch
f	frequency of nucleotide character
m	number of taxa
n	DNA sequence length
r	exchangeability in rate matrix
t	branch length
F	frequency of nucleotide character in bootstrap DNA sequence
D	discrete rate
H_0	null hypothesis of statistical test
I	invariate rate
N	frequency of the ML tree
K	number of clusters
Q	rate matrix
R	rejection rates of hypothesis
LBA	long branch attraction
ML(E)	maximum likelihood (estimates)
EL	expected log likelihood
MSEO	misspecification of exchangeability only
MSFO	misspecification of frequency only
MSRO	misspecification of rate only
JC69	Jukes-Cantor model
F81	Felsenstein 1981 model

Roman symbol	Description
EF	equal equilibrium frequency model
HKY	Hasegawa, Kishino and Yano model
GTR	generalized time-reversible model
GC	Goldman-Cox test
SSE	sum of square error
LRT	likelihood ratio test

Greek symbol	Description
Γ	among site rate variation
λ	constant rate of JC69
κ	transition/transversion ratio
π	equilibrium frequency of nucleotide character
τ	tree topology

ACKNOWLEDGEMENTS

I am extremly grateful to Dr. Hong Gu and Dr. Joseph Bielawski for their supervision on this thesis. I deeply appreciate for their helpful suggestions and many times of reviews on this thesis.

Special thanks go to Dr. Toby Kenney for his insightful discussions on the methodologies developed in this thesis.

I would like to thank Dr. Edward Susko and Dr. Toby Kenney for being the readers of this thesis.

Thanks to my friend, LinYun Ye and Joseph Mingrone for their computer software supports and Stuart Carson for his suggestion of modifications for this thesis.

This thesis would not have been finished without my girl friend CuiCui Wang's positive support.

CHAPTER 1

INTRODUCTION

In biology, phylogenetics is the formal name for the study for relationships between a set of various organisms. Cavalli-Sforza and Edwards (1967) indicated the phylogeny problem was actually a statistical inference problem. A phylogenetic tree is the structure to demonstrate the evolutionary history of life and it can be estimated from data having incomplete information (DNA sequences data) by using tree reconstruction methods.

There are various models of nucleotide substitution developed so far (Felsenstein 2004). However, due to the limitation of the information provided by some data and because statistical uncertainty is unavoidable, the estimated phylogenetic tree might not represent the true history of evolution. The tree reconstruction method can converge to a wrong phylogenetic tree if model assumptions are incorrect. This thesis first addresses the effect of model misspecification on one well known phylogenetic error, i.e., long branch attraction (Felsenstein 1978), by using simulation studies; then proposes and examines the performance of two goodness-of-fit tests for phylogenetic models of DNA sequences.

1.1 Phylogenetic tree and DNA sequence data

1.1.1 Phylogenetic tree

A phylogenetic tree contains (inner or external) nodes and branches. An m taxa phylogenetic tree is the representation of relationships among the m descendants (external nodes or tips) and unknown common ancestors (inner nodes). The branches are the connections between nodes. A topology refers to a branch order whereas a phylogenetic tree refers to both a branching order and a set of specified branch lengths. In general,

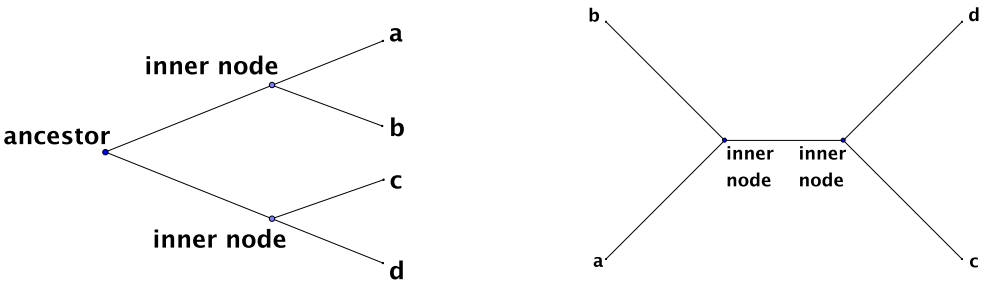
any tree topology can be rooted or unrooted (Fig 1.1). The total number of possible *m* taxa rooted tree topologies is:

$$1\ 3\ 5 \cdots (2m-3) = [(2m-3)!]/[2^{(m-2)}(m-2)!]$$

Thus, the total number of possible tree topologies increases geometrically as *m* increases. For the unrooted tree topologies, the total number of possible *m* taxa unrooted tree topologies is:

$$(2m-5)(2m-7)\cdots 5\ 3\ 1 = [(2m-5)!]/[(2^{(m-3)}(m-3)!)]$$

A tree topology can be represented graphically by drawing in a two-dimensional space or by plain text. For example, we can represent the tree topology in Fig 1.1 (a) in plain text as ((a,b),(c,d)). The text system makes different tree topologies easily distinguished and is more convenient for the computer-based research.



(a) Rooted Tree (b) Unrooted Tree

Figure 1.1: 4 taxa tree topology

1.1.2 DNA sequence data

A DNA sequence consists of 4 different nucleotide characters: adenine (A) and guanine (G) (Purines), cytosine (C) and thymine (T) (Pyrimidines). DNA sequence data typically include aligned DNA sequences. Each position of an alignment is called a site and a site pattern is the nucleotide characters in a particular site. Table 1.1 is an example of aligned DNA sequences with 4 taxa. The 1st site pattern is ACAA, and the second site pattern is AAAA.

	1	2	3	4	5	6	7	8	9	10
a	A	A	T	C	G	T	C	G	T	A
b	C	A	T	C	G	A	C	G	G	A
c	A	A	T	C	G	T	C	G	T	C
d	A	A	T	C	G	C	C	G	T	A

Table 1.1: Aligned DNA sequences with 4 taxa

The evolution of species can be considered as the consequence of the substitution of nucleotide characters in the DNA sequences of their ancestors. There are various continuous-time-Markov-process based statistical models for describing the changes of nucleotide characters among gene sequences.

1.2 Substitutions and relevant statistical models

1.2.1 Substitution matrix for evolution

For aligned DNA sequences, we assume the substitutions on each site are independent based on the same probabilistic model. If we start with a nucleotide character, say *i*, there are 4 possible changes: No change and the other three are the changes from *i* to other three nucleotide characters. Since *i* can be one of A, C, G, T, hence, there are $4 \times 4 = 16$ different ways of changes in total. A change is called a *transition* if it occurs within either pyrimidine or purine categories and is called a *transversion* if it occurs between a pyrimidine and a purine. Given that a change between nucleotide characters *i* and *j* occurs in a time interval *t*, the sum of probabilities of all possible changes equals to 1:

$$\sum_{j}^{A,C,G,T} p_{ij}(t) = 1 \tag{1.1}$$

Based on the equation (1.1), the substitution matrix is defined as:

	T	C	A	G
T	$p_{TT}(t)$	$p_{TC}(t)$	$p_{TA}(t)$	$p_{TG}(t)$
C	$p_{CT}(t)$	$p_{CC}(t)$	$p_{CA}(t)$	$p_{CG}(t)$
A	$p_{AT}(t)$	$p_{AC}(t)$	$p_{AA}(t)$	$p_{AG}(t)$
G	$p_{GT}(t)$	$p_{GC}(t)$	$p_{GA}(t)$	$p_{GG}(t)$

Table 1.2: Substitution matrix

Note that the summation of each row equals to 1.

1.2.2 Substitution rate matrix

Most molecular evolution models assume that a continuous-time Markov model along edges applies, which gives rise to the substitution matrix. The substitution matrix is in turn determined by the rate matrix, Q, for the process which is defined as the rate of change between nucleotide characters in an instant time *dt* is demonstrated in Table 1.3. If reversibility is assumed, the entries q_{ij} of matrix Q can be expressed by product of equilibrium frequencies of nucleotide characters $\pi = \{\pi_A, \pi_C, \pi_G, \pi_T\}$ and exchangeability $r = \{r_1, r_2, r_3, r_4, r_5, r_6\}$ (Table 1.4). The matrix Q of a substitution model depends on the particular settings of elements in π and R. In this thesis, I will employ four substitution models, JC69, F81, HKY and GTR, for studies.

	T	C	A	G
T	-	q_{TC}	q_{TA}	q_{TG}
C	q_{CT}	-	q_{CA}	q_{CG}
A	q_{AT}	q_{AC}	-	q_{AG}
G	q_{GT}	q_{GC}	q_{GA}	-

Table 1.3: A typical rate matrix Q

	T	C	A	G
T	-	$r_1\pi_C$	$r_2\pi_A$	$r_3\pi_G$
C	$r_1\pi_T$	-	$r_4\pi_A$	$r_5\pi_G$
A	$r_2\pi_T$	$r_3\pi_C$	-	$r_6\pi_G$
G	$r_4\pi_T$	$r_5\pi_C$	$r_6\pi_A$	-

Table 1.4: Q matrix used for modelling substitution

1.2.3 Substitution models

1.2.3.1 JC69 model

The JC69 (Jukes and Cantor 1969) is the simplest substitution model in phylogenetics because it assumes the exchangeabilities and character frequencies are all constant. Thus, the matrix Q is:

	T	C	A	G
T	-	λ	λ	λ
C	λ	-	λ	λ
A	λ	λ	-	λ
G	λ	λ	λ	-

Table 1.5: Q matrix of JC69

1.2.3.2 F81 model

The F81 (Felsenstein 1981) model is an extension of JC69, where the exchangeabilities are assumed to be 1 and character frequencies of A, C, G, T are not restricted ($\pi_A, \pi_C, \pi_G, \pi_T$). Thus, the entries of matrix Q depends on the base frequencies (Table 1.6)

	T	C	A	G
T	-	π_C	π_A	π_G
C	π_T	-	π_A	π_G
A	π_T	π_C	-	π_G
G	π_T	π_C	π_A	-

Table 1.6: Q matrix of F81

1.2.3.3 HKY model

The HKY (Hasegawa, Kishino and Yano 1985) model assumes the character frequencies are not restricted and the *Q* matrix depend on both character frequencies and *transition-transversion* ratio, denoted as κ. Hence, the *Q* matrix is:

	T	C	A	G
T	-	$\kappa\pi_C$	π_A	π_G
C	$\kappa\pi_T$	-	π_A	π_G
A	π_T	π_C	-	$\kappa\pi_G$
G	π_T	π_C	$\kappa\pi_A$	-

Table 1.7: *Q* matrix of HKY85

1.2.3.4 GTR (Generalized time-reversible) model

The GTR model (Lanave et al. 1984) is the most general DNA substitution model in phylogenetics. The GTR model depends on the character frequencies and the exchangeabilities, that is, they are fully defined by the model. Hence, the *Q* matrix is the form in Table 1.4 and it allows free parameters $\pi = (\pi_T, \pi_C, \pi_A, \pi_G)$ and exchangeability parameters $(r_1, r_2, r_3, r_4, r_5, r_6)$ with $r_f = 1$ fixed.

1.3 Tree reconstruction method: maximum likelihood (ML)

Given a substitution model and DNA sequence data, the maximum likelihood method can be employed for reconstructing the tree and estimating the branch lengths . Felsenstein (1981) presented the pruning algorithm to efficiently compute the likelihood function for a fixed set of sequences. Given aligned DNA sequences with *n* sites, substitution model and a tree topology τ, we want to estimate the branch lengths *t* (Fig 1.2) and the parameters in the substitution model. The transition probability of change from state *i* to state *j* on a branch of length *t* is denoted by $p_{ij}(t)$. Felsentein (1981) used two assumptions for calculating the likelihood (i) evolution in different sites is independent, and (ii) evolution in different lineages is conditionally independent, given their ancestral data.. The likelihood is calculated site by site. The general form of the likelihood (*L*) can be expressed by following:

$$L = p(\mathbf{X}|\Theta, \tau) = \prod_i p(\mathbf{X}_i|\Theta, \tau)$$

where X_i are the data at the *i*th site and Θ is a vector containing all branch lengths and parameters of the substitution model.

To demonstrate the likelihood calculation on a single site, I use the first site of Table 1.1 on the tips of a tree topology in Figure 1.2. The inner nodes e, g, f in Figure 1.2 are unknown ancestors. Based on the transition probability $p_{ij}(t_i)$ between nodes, the calculation of likelihood on this site is the summation over all possible states for the inner nodes:

$$L^1 = \sum_e \sum_g \sum_f \pi_g p_{ge}(t_1) p_{gf}(t_2) p_{eA}(t_3) p_{eC}(t_4) p_{fA}(t_5) p_{fA}(t_6)$$
$$e, g, f \in \{A, C, G, T\}$$

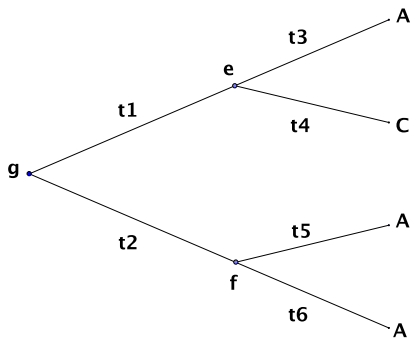


Figure 1.2: Tree for likelihood calculation, $t = (t_1, t_2, \cdots, t_6)$

Felsenstein (1981) gave an efficient algorithm for calculating this summation which would otherwise be prohibitive.

The likelihood of the observed DNA sequences in Table 1.1 is the product of the L^i's:

$$L = \prod_{i=1}^{10} L^i$$

Thus, the log likelihood score is the natural logarithm of L:

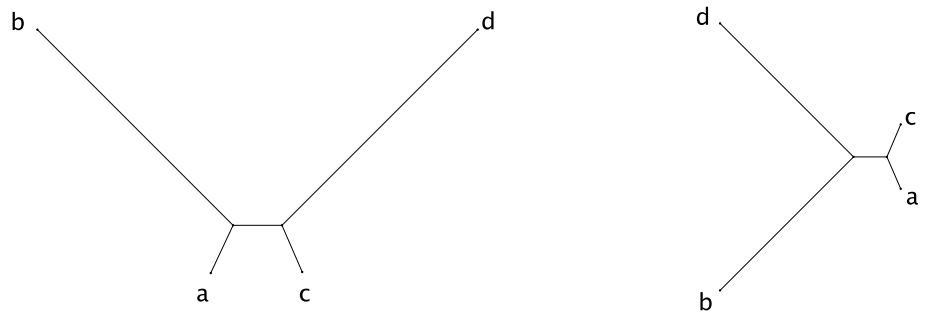
$$l = log(L) = log(\prod_{i=1}^{10} L^i) = \sum_{i=1}^{10} log(L^i)$$

Given a set of candidate tree topologies, the ML tree is the tree with highest likelihood score and it gives the highest probability of the data being observed. In phylogenetics,

the maximum likelihood approach sometimes cannot provide the correct estimate of tree when the true tree is in a relatively extreme case.

1.4 LBA phenomenon

If a tree has each of long terminal branches join with one of short branches and the branch between inner nodes is also relatively short, then, this kind of tree is referred to as a "hard tree" (Fig 1.3 (a)). The estimated tree based on data generated under a hard tree can mislead to a tree with the long branches grouping together, which is referred to as the LBA tree (Fig 1.3 (b)). This type of phylogenetic error is referred to as the long branch attraction (LBA), and the set of such hard tree topologies is known as the Felsenstein Zone. Bergsten (2005) reviewed several strategies to overcome the LBA issue: for example, by adding taxa to break up long branches, by improving evolutionary model or removal of fast-evolving species or genes.



(a) True tree (Hard) (b) LBA tree

Figure 1.3: Hard tree and LBA tree

In some simulation studies, the LBA phenomenon is not the only type of estimation bias under the ML method. Sometimes a type of bias called the long-branch-repel (Susko 2011) can occur under ML. However, this thesis will only focus on the LBA effect.

1.5 Summary of the thesis

This thesis explores the effects of misspecification and provides a goodness-of-fit test of substitution models. For both easy (a tree not in the Felsenstein Zone) and hard trees, Chapter 2 investigates the effects of the model misspecification based on simulation

studies. Both maximum likelihood and the *expected log likelihood* are employed in this chapter. Chapter 3 first discusses the model selection criteria and statistical tests used in phylogenetics. Then, two goodness-of-fit tests for testing the adequacy of substitution models are introduced. The sizes and powers of the proposed tests are demonstrated by simulation studies, which contain both small number and large number of taxa. In addition, some empirical data analyses are also included, and comparisons are made between the newly proposed test and the existing model test. Chapter 4 concludes the thesis.

CHAPTER 2

A SIMULATION STUDY OF THE EFFECT OF MODEL MISSPECIFICATION IN PHYLOGENETICS.

The ML method is a consistent method and it has been introduced for inferring phylogeny (Edwards and Cavalli-Sforza 1964, Felsenstein 1981). However, when the true tree is a hard tree (Fig 1.3 (a)) and the data are analyzed by a misspecified model, the maximum likelihood estimates (MLE) can be inaccurate estimates of the phylogenetic tree (Bruno and Halpern 1999, Brandon and Paul 1995, Heulsenbeck 1995).

The objective of this chapter is to employ simulation to investigate the measurable consequences of models with and without misspecification in inferring phylogenetics. To achieve this, a series of simulation studies are designed to target both 4 taxa easy and hard tree topologies. The hard tree is a known source that can result in phylogenetic error (see Chapter 1) and is known to depend on the level of model misspecification. Furthermore, another effect of model misspecification is investigated via the branch length estimates.

According to Kullback-Leibler information (Kullback and Leibler 1951), it is the *expected log likelihood* (EL) that should be maximized. The ML is a good approximation to the EL, thus, the EL can better reflect the results of ML method. In this chapter, I will use simulation to investigate the effect of different degrees of model misspecification on ML method for tree selection. Since the EL is also available, it will be used as a reference for the ML.

2.1 A review of Kullback-Leibler information

The Kullback-Leibler (KL) information is used for measuring the closeness of two probability distributions, and it serves as the basis of model selection criteria. Suppose a random variable $\mathbf{z}$ follows density function $f_\theta(\mathbf{z})$ and the density function $g_{\hat{\theta}}(\mathbf{z})$ is an approximation to density $f_\theta(\mathbf{z})$. The KL information, $I(\theta;\hat{\theta})$, is used to measure closeness of $f_\theta(\mathbf{z})$ and $g_{\hat{\theta}}(\mathbf{z})$. The best model for approximating $f_\theta(\mathbf{z})$ must have smallest KL information:

$$I(\theta;\hat{\theta}) = \int f_\theta(\mathbf{z}) \log f_\theta(\mathbf{z}) d\mathbf{z} - \int f_\theta(\mathbf{z}) \log g_{\hat{\theta}}(\mathbf{z}) d\mathbf{z} \tag{2.1}$$

$$= S(\theta;\theta) - S(\theta;\hat{\theta}) \tag{2.2}$$

$S(\theta;\theta)$ is a constant given $f_\theta(\mathbf{z})$, thus, $S(\theta;\hat{\theta})$ can be used to determine the goodness of fit of $g_{\hat{\theta}}(\mathbf{z})$. Then, to minimize KL information, we only need to maximize $S(\theta;\hat{\theta})$. $S(\theta;\hat{\theta})$ is referred to as the *expected log likelihood*. we can further express the second term of the right-hand side of equation (2.1) as:

$$\int f_\theta(\mathbf{z}) \log g_{\hat{\theta}}(\mathbf{z}) d\mathbf{z} = \int \log g_{\hat{\theta}}(\mathbf{z}) dF(\mathbf{z})$$

Suppose data $\mathbf{x} = (x_1, x_2, x_3, \cdots, x_n)$ are generated from an unknown distribution $F(\mathbf{x})$, and the maximum likelihood estimates $\hat{\theta}$ (MLE) can be estimated from $\mathbf{x}$. We can simply estimate the *expected log likelihood* by the *average log likelihood*:

$$\hat{S}(\theta,\hat{\theta}) = \frac{1}{n} \sum \log g_{\hat{\theta}}(x_i) \tag{2.3}$$

For a given substitution model, sequence data and a tree (τ) having *m* tips, the MLE can be easily computed. The number of unique site patterns is derived from all possible combinations of nucleotides (A, C, G, T) and is easily computed as $N = 4^m$. The matrix of unique site patterns ($\boldsymbol{X}$) is *N* by *m*, such that each row contains one of the set of possible site patterns. We denote the probabilities of all *n* site patterns under the true tree and true substitution model as $p_1^*, p_2^*, \cdots, p_N^*$.

If both the true probabilities (p^*'s) of site patterns in $\boldsymbol{X}$, and their probabilities under the MLE ($p(X_j|\hat{\theta},\tau)$, $j = 1, 2, \cdots, N$) for a given model are available, then the EL for

this model can be computed as below:

$$EL = \sum_{j=1}^{N} p_j^* log(p(X_j|\hat{\theta}, \tau)) \tag{2.4}$$

The true probabilities p^*'s are easily obtained by computing the exact probability of each site pattern given the true values of the parameters of the generating model and the true tree.

2.2 Simulation design

The design is comprehensive in covering the effects of three types of variables: (1) tree topologies with an associated set of branch lengths, (2) substitution models, and (3) sequence lengths. Because the true tree and the true generating model are known for each data set within the simulation study, it is possible to use the EL criterion as a reference of the ML criterion in tree selection.

A diagram of the simulation study is shown in Figure 2.1. I simulate and analyze sequence data under different scenarios, which are represented as format "generating model - analytical model". In general, the generating models are GTR, equal-equilibrium-frequency model (EF, nucleotide frequencies are each 0.25 and exchange abilities are free), and F81. The scenarios in this figure are GTR-JC69, GTR-EF, GTR-F81, F81-JC69 and EF-JC69, which represent different degrees of model misspecification.

In the generating models shown in Figure 2.1, all sites of the sequences are assumed to have evolved under the same rate ("equal rates" scenario), which is an unrealistic assumption for real data. Hence, I also generate sequences under GTR model but with the addition of discrete among site rate variation (ASRV), and this scenario is denoted as GTR+D. To simulate sequences under GTR+D model, for each data set, half the sites are simulated under one rate and the other half are simulated under another rate. The second rate for the heavy case is ten times the first rate and for the light case is twice the first rate.

The parameters of the GTR model are set to equal to the estimates from the globin-pseudogenes data (Yang 1994, Fig.1): $\hat{\pi}_T = 0.308$, $\hat{\pi}_C = 0.185$, $\hat{\pi}_A = 0.308$, $\hat{\pi}_G = 0.199$; $\hat{r}_1 = 0.987, \hat{r}_2 = 0.11, \hat{r}_3 = 0.218, \hat{r}_4 = 0.243, \hat{r}_5 = 0.395$. These estimates of exchangeabilities and equilibrium frequencies are also used for simulation under the EF

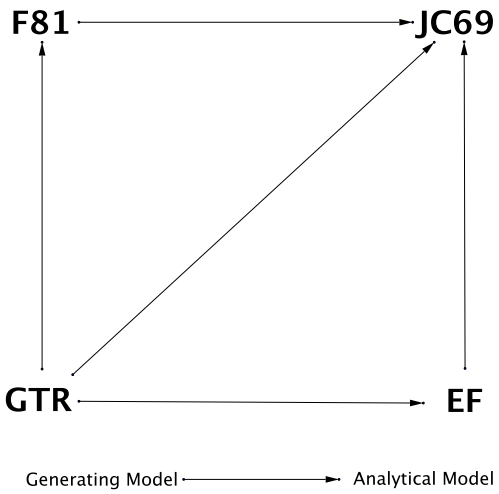
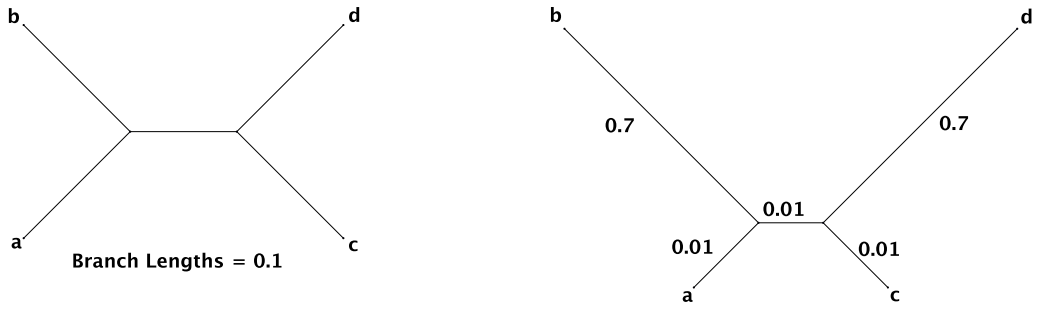


Figure 2.1: A diagram of simulation study

and the F81 model respectively.

The tree topologies for simulation contain easy and hard tree (Fig 2.2). Fig 2.2 (a) is the easy generating tree, the branch lengths are all set to 0.1. Fig 2.2 (b) is the hard generating tree, the short and long branch lengths are set to 0.01 and 0.7 respectively. Sequence lengths are 300, 500, and 1000 (and some scenarios contain 5000 or 10000) respectively. Based on these settings, *INDELible1.03* (Fletcher and Yang 2009) is used to generate 1000 replicates.



(a) Easy tree (b) Hard tree

Figure 2.2: Trees used for simulation

2.2.1 Notations

Let τ_1, τ_2 and τ_3 denote the 3 possible 4 taxa tree topologies, where $\tau_1 = \tau^*$ is the true tree. In the scenarios based on the hard tree, τ_2 is also referred as the LBA tree, and is denoted as τ_2^L. The frequencies that the ML tree is tree τ_1, τ_2 (τ_2^L), τ_3 are denoted as N_1, N_2 (N_2^L), and N_3 respectively.

2.3 Simulation study on data generated under easy tree

In this section, the experiment targets data that are generated under an easy tree and analyzed by either correct or incorrect (misspecified) models. In general, the scenarios contain GTR-GTR, JC69-JC69, GTR-JC69 and the heavy case of GTR+D-GTR. The GTR-JC69 has the highest degree of model misspecification because the GTR model is the most complicated generating model and the JC69 is the simplest analytical model. For the misspecification of ASRV, only the heavy case is selected for this study, since the heavy case is considered a higher degree of model misspecification than the easy case. The results of ML and EL for tree selection are below:

Sequence Length	Tree	Scenarios			
		GTR-GTR	JC69-JC69	GTR-JC69	GTR+D(heavy)-GTR
300	τ_1	1000	1000	1000	1000
	τ_2	0	0	0	0
	τ_3	0	0	0	0
500	τ_1	1000	1000	1000	1000
	τ_2	0	0	0	0
	τ_3	0	0	0	0
1000	τ_1	1000	1000	1000	1000
	τ_2	0	0	0	0
	τ_3	0	0	0	0

Table 2.1: Frequencies of estimated tree in 4 scenarios based on ML(EL) and easy simulation tree

The results (Table 2.1) show that the ML method converges to the correct tree (τ_1), even when the analytical model is different from the generating model. The N_1 is always 1000 for all scenarios and all sequence lengths. Hence, ML is a consistent method for data generated under an easy tree. The EL gives the same results as ML in all cases.

Thus, when the data are generated under the easy tree, ML is equivalent to EL for tree selection.

2.4 Simulation study on data generated under hard tree

In this section, data are generated under the hard tree, and analyzed under both correct and misspecified models.

For the cases with model misspecification, I start with the highest degree of model misspecification case, GTR-JC69, from which I can definitely observe LBA effect. I then explore the impact of reduced degree of model misspecification through different restrictions on the parameters of the substitution model. The scenarios contain: (1) GTR-F81, (2) F81-JC69, (3) GTR-EF, and (4) EF-JC69. These scenarios have reduced degree of model misspecification through the use of the two intermediate models (F81, EF) (Fig 2.1). In addition, both heavy and light cases of the GTR+D-GTR are also included for comparison.

Among the scenarios above, another interesting question is whether the LBA effect only impacts the estimate of topology. To investigate this problem, for those data whose tree topologies are correctly estimated, the branch length estimates are compared with the true branch lengths.

2.4.1 The performance of ML and EL in tree selection without model misspecification

The results that the data are generated using a hard tree and analyzed without model misspecification are shown in Table 2.2. The performance of the ML is not as good as for the data simulated under the easy tree. In the GTR-GTR case, when the sequence length is 300, N_1 is 568 whereas N_2^L is 269 and N_3 is 163. N_1 increases to 676 and 794 as the sequence length increases to 500 and 1000 respectively. In the JC69-JC69, N_1 increases from 586 to 807 when the sequence length increases from 300 to 1000. The results are clearer under the EL. Under the GTR-GTR, N_1 is 811 under EL and it increases to 919 when the sequence length increases to 1000. In the JC69-JC69, N_1 is 831, 889 and 935 when the sequence length is 300, 500 and 1000 respectively. Hence, ML can also converges to the true tree as the sequence length becomes longer if the

analytical model is correct.

Sequence Length	Tree	Scenarios Based on ML		Scenarios Based on EL	
		GTR-GTR	JC69-JC69	GTR-GTR	JC69-JC69
300	τ_1	568	586	811	831
	τ_2^L	269	250	85	69
	τ_3	163	161	104	100
500	τ_1	676	645	857	889
	τ_2^L	207	224	73	49
	τ_3	117	130	70	62
1000	τ_1	794	807	919	935
	τ_2^L	125	115	54	30
	τ_3	81	78	27	35

Table 2.2: Frequencis of trees estimated based on ML, EL and hard tree in 1000 simulations

2.4.2 The performance of ML and EL in tree selection with model misspecification

In this section, the simulation studies are based on the sequence data which are simulated under the hard tree and analyzed under misspecified model, hence, there is a very strong possibility that results could be led to the LBA phenomenon. The negative impact on phylogenetic inference is easily observed in a scenario with "heavy" model misspecification, i.e. GTR-JC69. For different sequence lengths in GTR-JC69, the ML tree is most often τ_2^L, which is the LBA tree. This result confirms the classic LBA estimation error. ML is converging to the LBA tree as sequence length increases from 300 to 1000 (Table 2.3). Note that model misspecification is considered "heavy" here because both the DNA exchangeability parameters and the equilibrium frequencies are misspecified.

The performance of EL in the GTR-JC69 indicates that EL converges to the LBA tree at a faster rate than ML as sequence length increases. In the GTR-JC69, N_2^L under EL increases from 724 to 898 as sequence length increases from 300 to 1000. This supports the notion that GTR-JC69 is the case with the most serious model misspecification.

For both scenarios with misspecification of exchangeabilities only (MSEO: GTR-F81 and EF-JC69), ML is converging to the LBA tree as sequence length increases, but not as quickly as in GTR-JC69 (Table 2.4). In GTR-F81, N_2^L under ML is 546, 592, and 620

		Sequence Lengths								
		300			500			1000		
Scenario	Method	τ_1	τ_2^L	τ_3	τ_1	τ_2^L	τ_3	τ_1	τ_2^L	τ_3
GTR-JC69	ML	287	626	85	307	640	53	312	666	22
	EL	260	724	16	152	844	4	100	898	2

Table 2.3: Frequencies of trees estimated in 1000 simulations based on ML and EL and hard tree

when sequence lengths are 300, 500, and 1000 respectively. Convergence to the LBA tree is slower compared to GTR-JC69, because the degree of model misspecification is lower. Interestingly, in GTR-F81, N_1 values are similar for the sequence lengths 300, 500, and 1000 (346, 345, 352, respectively). Although ML also converges to the LBA tree under EF-JC69, the rate is even slower than under GTR-F81. Indeed, it is difficult to draw a firm conclusion for this case from sequences of 1000 sites, so additional simulation is carried out for 5000 sites. This last simulation shows the LBA artifact, as N_2^L is 551, and N_1 is 448. These results highlight the interaction between exchangeabilities and frequencies with respect to the convergence to the LBA tree; when equilibrium frequencies are unequal (as is typically the case with real data), the inadequate modeling of exchangeabilities has a bigger impact on the convergence rate to the LBA tree.

	Sequence Lengths											
	300			500			1000			5000		
MSEO Scenarios	τ_1	τ_2^L	τ_3	τ_1	τ_2^L	τ_3	τ_1	τ_2^L	τ_3	τ_1	τ_2^L	τ_3
GTR-F81	346	546	108	345	592	63	352	620	28	NA	NA	NA
EF-JC69	416	459	125	421	499	80	469	493	38	448	551	1

Table 2.4: Frequencies of tree estimated in 1000 simulations based on ML and hard tree

The EL better presents the LBA effect for these two cases as sequence length increases (Table 2.5). In the GTR-F81 case, N_2^L under EL increases from 628 to 904 as sequence length increases from 300 to 1000. In the EF-JC69 case, the results of EL demonstrate that N_2^L under EL increases from 434 to 645 as sequence length increases from 300 to

1000.

	Sequence Lengths								
	300			500			1000		
MSEO Scenarios	τ_1	τ_2^L	τ_3	τ_1	τ_2^L	τ_3	τ_1	τ_2^L	τ_3
GTR-F81	337	628	35	242	749	9	96	904	0
EF-JC69	511	434	55	473	503	24	354	645	1

Table 2.5: Frequencies of trees estimated in 1000 simulations based on EL and hard tree

For those cases with misspecification of frequencies only (MSFO), convergence to the LBA tree under ML is very slow (GTR-EF) or non-existent (F81-JC69) (Table 2.6). Under GTR-EF, N_1 under ML is 461, 528, and 505 when sequence length is 300, 500, and 1000 respectively. Only after increasing sequence length to 10000 dose the ML begin to clearly favor the LBA tree since N_2 is 627. The EL provides clearer results, as sequence length increases, N_2^L under EL increases from 214 to 369 while N_1 under EL decreases from 559 to 319 (Table 2.7). This pattern indicates that EL favors τ_2^L and will converge to τ_2^L eventually. These results are different for F81-JC69 case. In the F81-JC69, N_1 under ML is most frequent (535) when sequence length is just 300, and this number increases to 723 when sequence length is 1000. Under the EL, N_1 is 817, 843, 915 as sequence length increases from 300 to 1000. Both the ML and EL of F81-JC69 confirm that there was no LBA in this case. The reason for this difference is the exchangeabilities are different under the GTR and EF models ($r_1 = 0.987, r_2 = 0.11, r_3 = 0.218, r_4 = 0.243, r_5 = 0.395, r_6 = 1$), whereas under the F81 and JC69 models, the exchangeabilities are all equal to 1. For both of the generating models (GTR and F81), the true frequencies are 0.308, 0.185, 0.308 and 0.199; they are misspecified in the analytical models (EF and JC69) by setting all of them equal to 0.25. The degree of misspecification is relatively mild (i.e., 0.25 is not very different from any of the true values). The lack of LBA effect in the F81-JC69 case, and its occurrence in the GTR-EF case, suggests that the impact of frequency misspecification is context dependent.

The misspecification of among site rate variation (MSRO) is investigated for both a heavy case and a light case, according to the ratios of the tree lengths for two categories

	Sequence Lengths											
	300			500			1000			10000		
MSFO Scenarios	τ_1	τ_2^L	τ_3	τ_1	τ_2^L	τ_3	τ_1	τ_2^L	τ_3	τ_1	τ_2^L	τ_3
GTR-EF	461	282	157	528	256	134	505	277	107	221	627	2
F81-JC69	535	315	149	648	242	107	723	218	59	NA	NA	NA

Table 2.6: Frequencies of tree estimated in 1000 simulations based on ML and hard tree

	Sequence Lengths								
	300			500			1000		
MSFO Scenarios	τ_1	τ_2^L	τ_3	τ_1	τ_2^L	τ_3	τ_1	τ_2^L	τ_3
GTR-EF	559	214	227	467	284	249	319	369	312
F81-JC69	817	95	88	843	94	63	915	61	24

Table 2.7: Frequencies of tree estimated in 1000 simulations based on EL and hard tree

of sites within a sequence (heavy case: 1:10, light case: 1:2) (Table 2.8). Under the light case, the results demonstrate that N_1 is the largest and the misspecification of rate does not impact the tree selection. Specifically, N_1 under ML is 480, 524, 571 when sequence length is 300, 500, and 1000 respectively. Results for the heavy case are the opposite of those for the light case in that there is an LBA effect. When sequence length is 300, N_2^L for ML is the largest (691), which is the LBA tree. This number increases to 774 and 853 respectively when the sequence length increases to 500 and 1000. Thus, incorrect modeling of among site rate variation can contribute to LBA when rate differences among sites are large enough.

The results under EL make the above points clearer (Table 2.9). In the light case (LBA absent), N_1 under EL is 716, 769 and 882 for the data with sequence length of 300, 500 and 1000 respectively. For the heavy case (LBA present), N_2^L under EL is 835, 928, 996 for the data with sequence length of 300, 500, and 1000 respectively.

Thus LBA effect can impact the tree selection when the sequences are generated under a hard tree and analyzed with misspecified models and the percentage of LBA effect

	Sequence Lengths								
	300			500			1000		
MSRO Scenarios	τ_1	τ_2^L	τ_3	τ_1	τ_2^L	τ_3	τ_1	τ_2^L	τ_3
GTR+D(light)-GTR	480	431	89	524	407	69	571	394	35
GTR+D(heavy)-GTR	241	691	68	189	774	37	140	853	7

Table 2.8: Frequencies of tree estimated in 1000 simulations based on ML and hard tree

	Sequence Lengths								
	300			500			1000		
MSRO Scenarios	τ_1	τ_2^L	τ_3	τ_1	τ_2^L	τ_3	τ_1	τ_2^L	τ_3
GTR+D(light)-GTR	716	153	131	769	131	100	882	64	54
GTR+D(heavy)-GTR	151	835	14	69	928	3	4	996	0

Table 2.9: Frequencies of tree estimated in 1000 simulations based on EL and hard tree

depends on the details of how the analytical model is misspecified. In next section, the effect of model misspecification on the parameter estimates will be investigated.

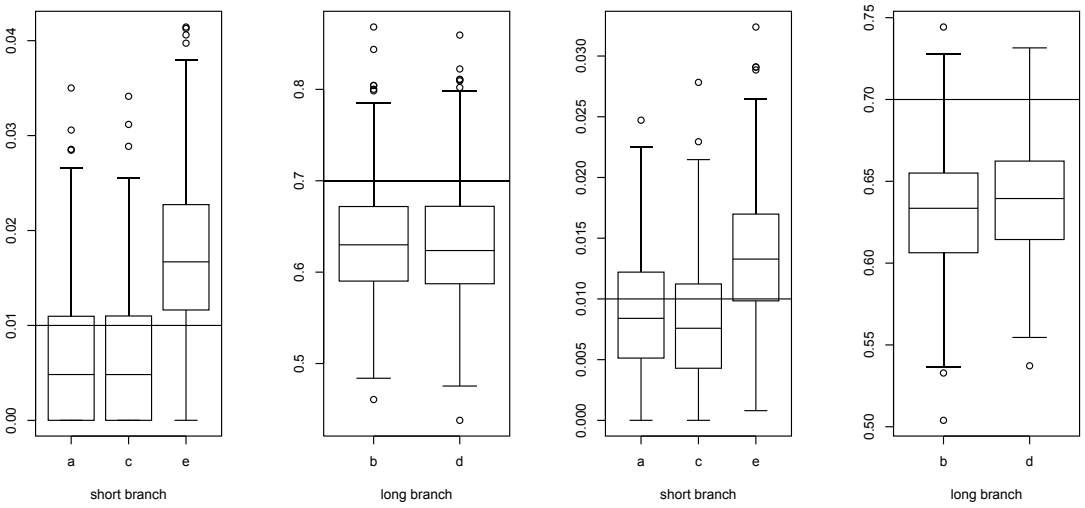
2.4.3 Branch length estimates are impacted by model misspecification

This section investigates the effect of model misspecifications other than those already documented for the tree. To do this I investigate branch lengths estimated for the true tree (i.e., where there is no LBA artifact). The ML estimates of branch-lengths are presented in box-plots where the short branches are denoted as *a*, *c*, and *e* and the long branches are denoted as *b* and *d*. To completely remove the impact of LBA, the results for these plots are restricted only to those replicates where ML correctly selected τ_1 as the ML tree despite the misspecification of the analytical model. The reason that I only consider these results is because the optimized branch length parameters on τ_1 can be compared with the true parameters (i.e., branch length parameters of the true tree). In this sense, these plots demonstrate the departure of the branch estimates from the true values, and I expect to see some relationship between the branch length estimates and

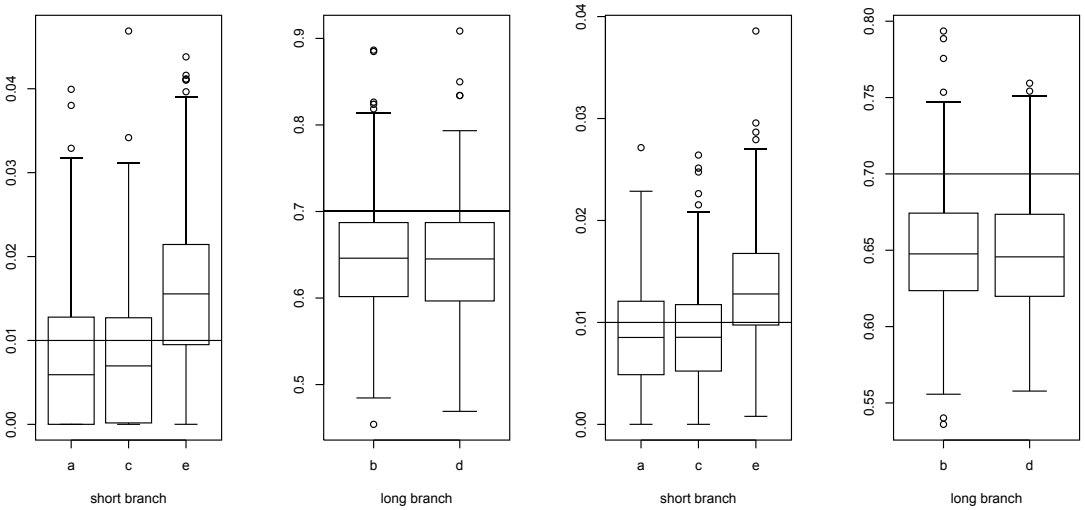
model misspecification.

In the MSEO (Fig 2.3) and GTR-JC69 (Fig 2.4) cases, the median of the estimates of the longest branch are lower than the true value, indicating a tendency towards underestimation of this type of branch. The true values are out of the "box", which means they are over the third quartile; this indicates that even when the tree is correct, the effect of model misspecification can be substantial. For the short branch estimates, branch *a* and *c* are also underestimated, but branch *e* (internal branch) is over-estimated. When the sequence length increases to 1000, the estimates of branch lengths do not improve; the estimates still depart from the true values. However, the number of outliers is reduced for most branches.

Recall that the LBA-inconsistency is absent from the F81-JC69 case. Hence, it is not surprising that the branch length estimates for τ_1 are better than in the MSFO cases (Fig 2.5). Specifically, the long-branch estimates are close to the true values, and the short branch estimates all lie in the "box" (Fig 2.5 (a)). When the sequence length increases to 1000, estimates under the F81-JC69 case further improve (Fig 2.5 (b)) with all of the branch estimates close to the true values. Interestingly, the long-branch estimates under GTR-EF (Fig 2.5 (d)) with sequence length of 1000 have a similar appearance to F81-JC69, but the short branch estimates are different because some of them are out of the "box" under GTR-EF (Fig 2.5 (d)). When the sequence length increases to 10000, even the long-branch estimates begin to depart from the true values, and the estimates of short branches exhibit much larger divergence (Fig 2.5 (e)). These results indicate that negative effects of model misspecification are not restricted to just the tree, as the estimates of branch lengths under the true tree are unsatisfactory.

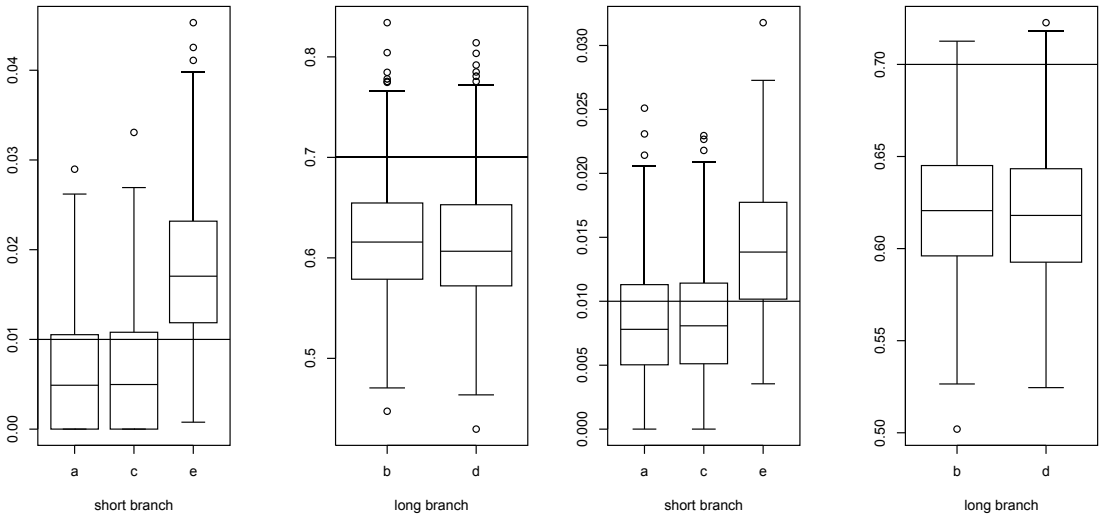


(a) GTR-F81 length=300 (346 trees) (b) GTR-F81 lenth=1000 (352 trees)



(c) EF-JC length=300 (416 trees) (d) EF-JC length=1000 (469 trees)

Figure 2.3: Branch estimates of MSEO, the dash lines represent the true branch lengths



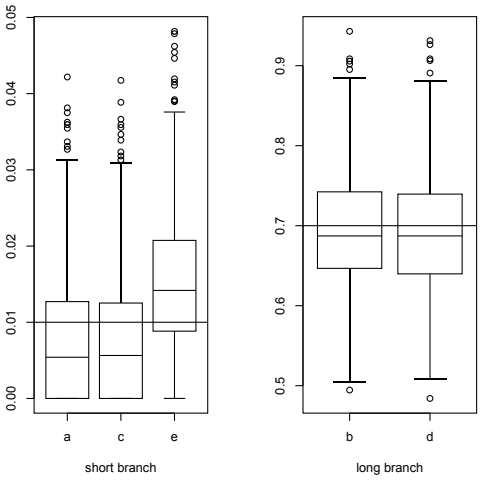
(a) GTR-JC69 length=300 (287 trees) (b) GTR-JC69 length=1000 (312 trees)

Figure 2.4: Branch estimates of GTR-JC69, the dash lines represent true branch lengths

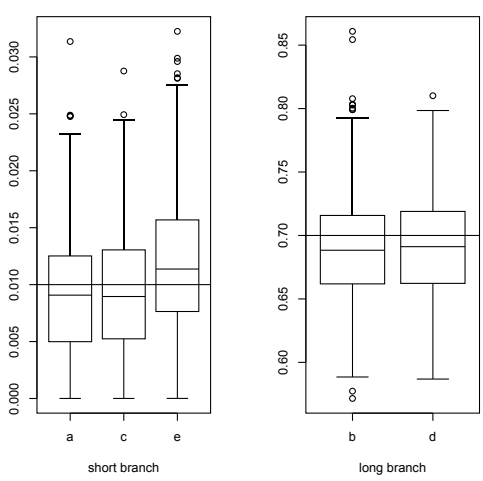
Branch estimates are also investigated for both light and heavy cases of MSRO. Recall that the ML under light MSRO does not converge to the LBA tree. Hence, the box-plots (Fig 2.6) indicate that the branch lengths are somewhat well estimated, because the true value lines do not depart away from the "box" too much when sequence lengths are 300 as well as 1000. For heavy MSRO (Fig 2.7), where the ML converges to the LBA tree, the box-plots are different. When the sequence length is 300, the true value of the long-branch is far from the "box", with most of long branches being seriously underestimated. When the sequence length is 1000, estimates of long branches are even worse, because there is absolutely no overlap between the estimates and the true values. The short branch estimates are not as impacted as the long branch; in general, they do not depart from the true values except the internal branch, e. Taken together with previous results, this investigation indicates that the misspecification of any aspect of the model (exchangeabilities, frequencies, or rates among sites) can have impact beyond the tree.

2.5 Conclusion

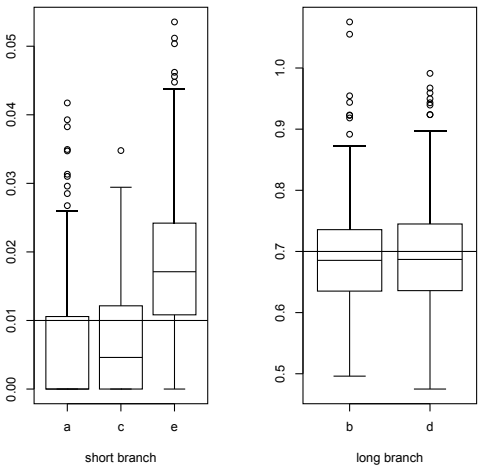
Both the ML and the EL perform very well for the data generated from an easy tree, which is the easy estimation problem, regardless of the analytical model. For the data



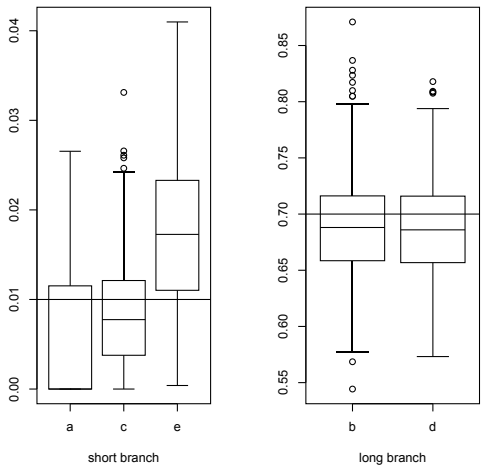
(a) F81-JC69 length=300 (817 trees)



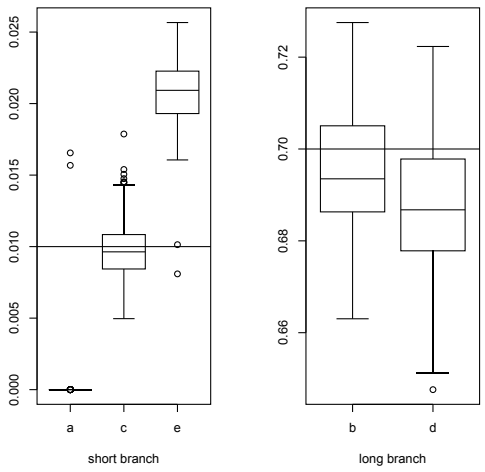
(b) F81-JC69 length=1000 (915 trees)



(c) GTR-EF length=300 (461 trees)

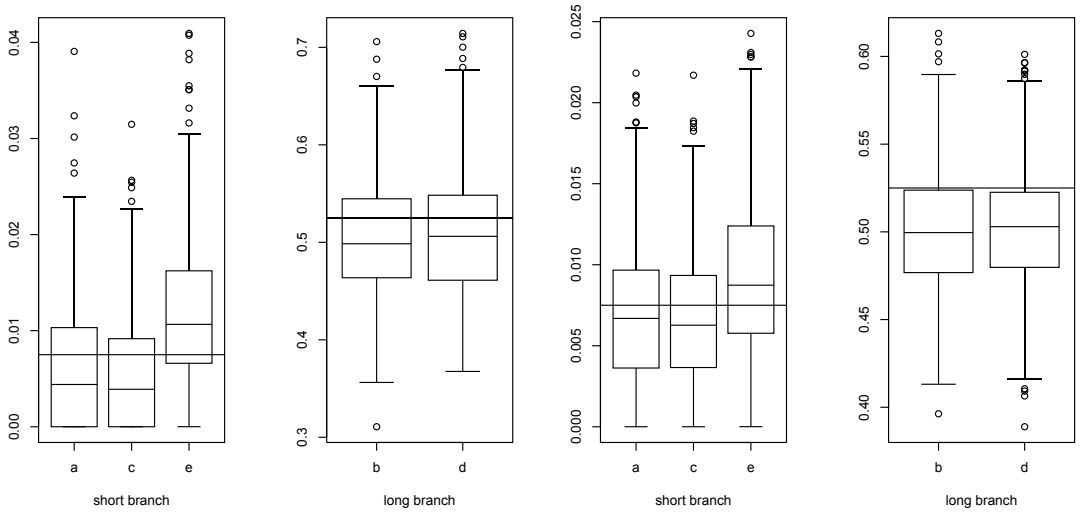


(d) GTR-EF length=1000 (505 trees)



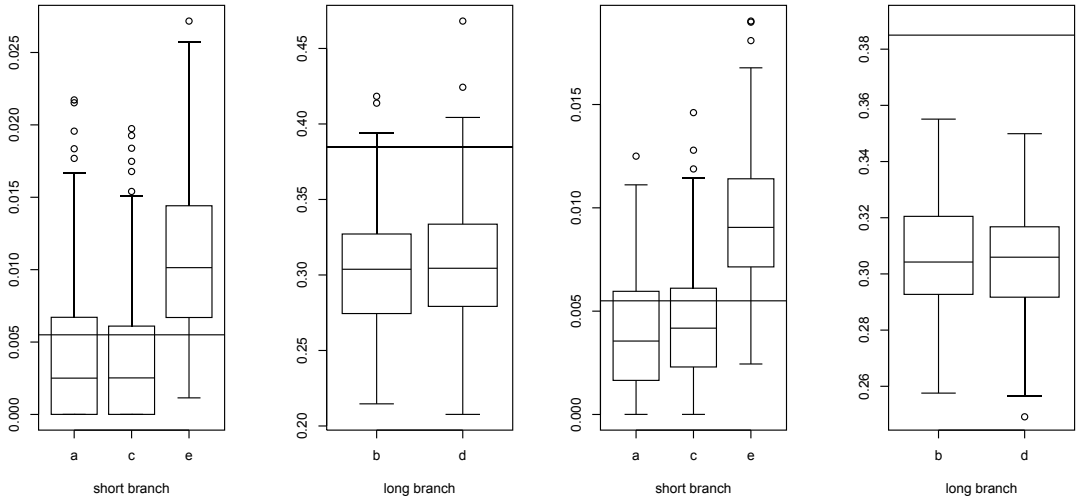
(e) GTR-EF length=10000 (221 trees)

Figure 2.5: Branch estimate of MSFO, the dash lines represent the true branch lengths



(a) GTR+D(light)-GTR length=300 (480 trees) (b) GTR+D(light)-GTR length=1000 (571 trees)

Figure 2.6: Branch Estimate of GTR+D(light)-GTR, dash lines represent true branch lengths



(a) GTR+D(heavy)-GTR length=300 (241 trees) (b) GTR+D(heavy)-GTR length=1000 (140 trees)

Figure 2.7: Branch Estimate of GTR+D(heavy)-GTR, the dash lines represent the true branch lengths

generated under hard tree, ML can converge to the true tree if the analytical model is correct. But if the sequence length is insufficient, the ML does not necessarily estimate the true tree due to the large variance of the estimate. In this case, as a reference of ML, the EL can better confirm the direction of convergence even for the data with short sequence length. For the data generated under hard tree and analyzed with misspecified models, the ML trees converge to the LBA tree except in some cases with relatively slight misspecification (light MSRO case and F81-JC69). For the cases with mild model misspecification (GTR-EF, EF-JC69), ML cannot give a clear conclusion unless an extremely long data sequence is available. In every case, EL always gives a much more clear conclusion than ML. The impact of model misspecification is not only on the estimate of the tree. Within the LBA scenarios, even when the ML tree is the true tree, the branch length are often very biased.

From these simulations, I find the ML cannot correctly estimate the true tree in the cases of model misspecification for data generated under hard tree. Thus it is necessary to select an adequate model for analysis. I will look into the model adequacy tests in the next chapter.

CHAPTER 3

GOODNESS-OF-FIT TESTS FOR ADEQUACY OF DNA SUBSTITUTION MODELS

It is important to select an adequate substitution model since an inadequate model can negatively impact phylogenetic inference. There are several methods which are based on information theory criteria and statistical tests already developed for this purpose. Posada and Crandall (1998) have developed software to test the adequacy of a substitution model by using method based on ML score. Waddell et al. (2008) presented the statistical test based on binning site patterns and maximum likelihood ratio test. In this chapter, two more novel methods based on binning of site patterns and Pearson's χ^2 goodness-of-fit test will be developed for testing the adequacy of substitution models.

3.1 A review of model selection criteria in phylogenetics

There are several formal criteria that have been employed for model selection in the phylogenetic context. Sullivan and Joyce (2005) summarized 4 criteria: hierarchical likelihood ratio test (hLRT), Akaike's information criterion (AIC), Bayesian model selection (BIC) and the decision theory (DT). All of these methods select the best model from a set of candidate models. Thus, the relatively better model is selected. A well selected set of candidate models is necessary for these methods, but this requirement is difficult to meet in reality.

Goldman (1993) developed a method for testing the adequacy of a phylogenetic model

based on Cox text and Monte-Carlo simulation, referred to as the Goldman-Cox test (GC test). In the context of phylogenetics, the sequence data contain both the information provided by the tree and the substitution model. The null hypothesis of the GC test is composite:

H_0: (a) the sequences are related by an unknown phylogenetic "tree" structure

(b) the sites of sequences have evolved independently, according to the specific model.

Thus, a desired alternative hypothesis is to assess both (a) and (b) in the null hypothesis. Goldman (1993) considered an unrestricted alternative hypothesis, that the sites are independently and identically distributed and the probability that each site exhibits a particular pattern *s* is *p(s)*. Let *S* be a set that contains 4^m site patterns under an *m* taxa tree and *n* is the length of DNA sequences, then the alternative hypothesis is:

H_a: probability of site *i* exhibits pattern $s \in S$ is $p(s)$, $\forall$ i=1,2 $\cdots$ n.

For a given sequence data, the test statistic of the LRT based on the null and the alternative hypotheses is:

$$\hat{\delta}_D = \hat{l}_a - \hat{l}_0 \tag{3.1}$$

where $\hat{l}_0$ is the maximum log likelihood of the sequence data under H_0. Let n_s denote the number of sites exhibiting the site pattern *s*, then, the likelihood function under H_a is:

$$L_a = \prod_{s \in S} p(s)^{n_s}$$

The maximum likelihood estimate (MLE) for $p(s)$ is then:

$$\hat{p}(s) = \frac{n_s}{n}$$

The maximized log likelihood under H_a is simply:

$$\begin{aligned}\hat{l}_a &= log[\prod_{s\in S}(\frac{n_s}{n})^{n_s}] \\ &= \sum_{s\in S} n_s log(n_s) - nlog(n)\end{aligned}$$

In principle, the likelihood ratio statistic has an approximate χ^2 distribution with degrees of freedom equal to the number of patterns minus the number of estimated parameters in the model. In practice, because the number of possible patterns is usually large, this approximation does not work well. To assess the null hypothesis, Goldman employed a parametric bootstrap to simulate a set of sequences based on the MLEs of the original data under the null hypothesis. For each simulated data set, the test statistic is calculated according to equation (3.1) and they form the null distribution. The $\hat{\delta}_D$ is then compared to the null distribution. If $\hat{\delta}_D$ is larger than the 95th percentile in the null distribution, then H_0 is rejected at 5% level.

Ripplinger and Sullivan (2010) compared the simplest models, which are not rejected by the GC test and the Bayesian posterior predictive simulations (PPS) for testing the adequacy of various substitution models, with the models selected by model selection criteria (hLRT AIC, BIC, DT). These studies were based on empirical data and simulations. The results demonstrated that the GC test and the PPS normally selected simpler models than those selected by model selection criteria. The PPS failed to reject the simpler models selected by the GC test and the GC test mostly failed to reject the simple models incorporating among site variation, such as JC+$I + \Gamma$. The model selection criteria only select the relatively better models from the alternative models. The GC test employed the multinomial distribution of site patterns, which is more general than any substitution models. Thus, it can be used as a goodness-of-fit test between a model and the sequence data. However, the parametric bootstrap employed by the GC test makes this test computationally expensive, especially for large number of taxa.

3.2 Pearson's χ^2 test

As a goodness-of-fit test, the Pearson's χ^2 test compares the observed frequency distribution and the expected frequency distribution under the null hypothesis for categorical

data. The test statistic is:

$$\chi_o^2 = \sum_i^K \frac{(O_i - E_i)^2}{E_i} \tag{3.2}$$

where the O_i and E_i are the observed frequency and the expected frequency of *i*th category and the K is the number of categories. The test statistic follows the χ^2 distribution with degree of freedom $K - 1$.

Goldman (1993) indicated that the Pearson's χ^2 test can not be used in the context of phylogentics because the size of each category, which is referred to as each pattern, in goodness-of-fit test is required to be at least five or the sample size should be at least four or five times the number of categories. These requirements are difficult to meet in phylogenetics because some patterns appear very rarely under particular models and the number of site patterns will increase rapidly as the number of taxa increases.

In this Chapter, two goodness-of-fit tests are developed to overcome these issues by binning site patterns. The first test bins the site patterns based on their supports to different tree topologies and the second test bins the site patterns based on their character frequencies. After binning the site patterns, the sequence data are rearranged into several bins and the bins serve as the categories in a multinomial distribution. Thus, the χ^2 distribution can be used since the aforementioned conditions no longer exist. These two tests are applied in two different situations. The size and power of the tests are demonstrated by simulation.

3.3 A goodness-of-fit test for testing both the tree and the substitution model

This section will demonstrate a site pattern binning method using an example of 4 taxa tree, to test:

H_0^j: (a) the tree τ_j is true, (b) the analytical model *M* is the true model.

Note that the null hypothesis is indexed by the *j*th tree, the assumption being the substitution model is the same for each tree being tested.

3.3.1 Binning method

Based on parsimony (Edwards and Cavalli-Sforza 1963), a site pattern with the smallest number of changes under tree τ has the maximum parsimony score and it supports the τ. For 4 species, there are 256 different site patterns in total that could be possibly observed among the *n* sites. The binning of 256 site patterns is based on their support for different tree topologies with the support calculated by parsimony:

Binning procedures

- Non-informative bin: This bin includes the constant pattern xxxx (e.g., AAAA), or the singleton patterns xxxy, xxyx, xyxx and yxxx (e.g., GCCC, AAAT, CCTC,$\cdots$) and the patterns contain 4 different nucleotide characters. None of these sites contain signal for any tree, under parsimony, thus give the same support to any tree .

- Informative bins: Some patterns contain strong signals for the history of substitution that correspond to a particular tree . For example, pattern xxyy (e.g., AACC, AAGG, CCTT, $\cdots$) strongly supports tree (1,2),(3,4). The xxyy can be the consequence of only one substitution between the inner node 5 and 6 (Fig 3.1). Hence, these patterns are binned to the same bin. Similarly, the patterns xyxy and xyyx are binned into two other bins since they strongly support trees (1,3),(2,4) and (1,4),(2,3) respectively.

- Semi-informative bins: Site patterns, xxyz (e.g., AACT) and yzxx (e.g., GACC) support tree (1,2),(3,4), but less strongly, Since it needs more than one substitution from the inner node to the tips (Fig 3.2). Hence, these are binned into the same bin. Similarly, the site patterns xyxz, yxzx are binned to the same bin and the site patterns xyzx, zxxy are binned to the same bin.

Thus, there are 7 bins in total:

(1) XXYY
(2) XYXY
(3) XYYX
(4) XXYZ, YZXX
(5) XYXZ, YXZX

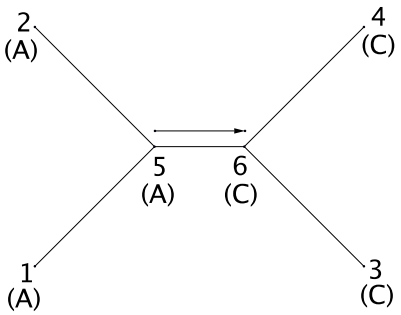


Figure 3.1: One substitution A $\rightarrow$ C

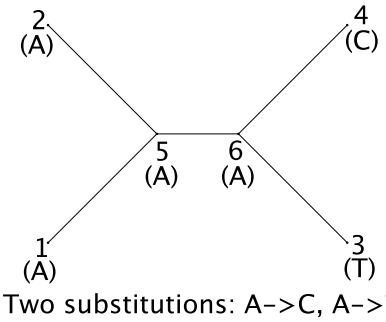


Figure 3.2: Two substitutions assuming both inner nodes are A

(6) XYZX, YXXZ

(7) Constant + singleton patterns+XYZW

Since the definition of bins is based on the site support to different topologies, the frequency distribution over these bins should be more sensitive to the tree than the substitution model in the null hypothesis.

Based on the bins defined above, I calculate the test statistic in (3.2) and perform the goodness-of-fit test according to the following procedure:

1. Compute the observed count of each bin O_i from the sequence data.

2. Compute the MLE for the model parameters (branch lengths and substitution model parameters) under the null hypotheses.

3. Calculate the expected probability of the 256 site patterns using the MLE in step 2.

4. Compute the expected probability of the *i*th bin P_i as sum of the expected probabilities of all the site patterns in the bin.

5. The expected count of the *i*th bin is $E_i = nP_i$, where *n* is the sequence length.

6. The goodness-of-fit test statistics X_o^2 can be easily obtained:

$$X_o^2 = \sum_{i=1}^{7} \frac{(O_i - nP_i)^2}{nP_i}$$

7. Compare to χ^2 distribution with $df = 6$ to conclude the test.

The steps 2-7 are repeated for each tree τ_1, τ_2, and τ_3.

3.3.2 Simulation design

200 data sets are simulated with sequence lengths equal to 300, 500, 1000, and 10000 for each combination of generating model and tree . The generating models include GTR and GTR+D. The tree topologies include an easy tree and a hard tree (Fig 2.2). The parameters of the GTR model are: $\pi_T = 0.308$, $\pi_C = 0.185$, $\pi_A = 0.308$, $\pi_G = 0.199$; $r_1 = 0.987, r_2 = 0.11, r_3 = 0.218, r_4 = 0.243, r_5 = 0.395, r_6 = 1$. The GTR+$D$ contains both heavy and light scenarios. The second rate for the heavy case is ten times the first rate and for the light case is twice the first rate. *INDELible1.03* is used for simulation. The scenarios in this simulation study contain the cases with different degrees of model misspecification: GTR-GTR, GTR-JC69, GTR-HKY, GTR-F81, GTR+D-GTR.

3.3.3 Analysis results

3.3.3.1 The size and power of the test when the true tree is easy

The results of the statistical test when true tree is an easy tree depend on different combinations of the models used for simulation and analysis. In the scenarios without model misspecification, referred to as the GTR-GTR in Table 3.1, the test is actually testing for the tree . The results for different sequence lengths (300, 500, 1000, 10000) are similar. The rejection rates under the null hypothesis H_0^2 and H_0^3 (the tree in the null hypothesis is wrong) for different sequence lengths are all 100%. This indicates that when the null model is correct, the hypotheses of wrong trees can be rejected. The size of the test can be obtained when both the tree and model in null hypothesis are correct. They are around 5% and are satisfactory.

		Scenarios		
Seqence Lengths	Hypothesis	GTR-GTR	GTR-JC69	GTR-F81
300	H_0^1	6%	17%	14%
	H_0^2	100%	100%	100%
	H_0^3	100%	100%	100%
500	H_0^1	5%	22%	19%
	H_0^2	100%	100%	100%
	H_0^3	100%	100%	100%
1000	H_0^1	5.5%	42.5%	33.5%
	H_0^2	100%	100%	100%
	H_0^3	100%	100%	100%
10000	H_0^1	5%	100%	100%
	H_0^2	100%	100%	100%
	H_0^3	100%	100%	100%

Table 3.1: Rejection rates for each hypothesis in 3 analysis model scenarios when true tree is an easy tree

For the GTR-JC69 (Table 3.1), which has a high degree of model misspecification, the rejection rates under the H_0^2 and H_0^3 are also 100% regardless of sequence lengths. The power of the test when the tree is true but the substitution model is wrong is relatively low when the sequence length is small. When the sequence length is sufficiently long, the power can be 100%. This indicates that when the null model is wrong, the parameters of the substitution model are estimated incorrectly. And the expected count of each bin can be impacted based on the poor estimates. For a mild model misspecification case, which is referred to as GTR-F81 (see Chapter 2) (Table 3.1), the rejection rates under H_0^1 with the sequence lengths less than 10000 (14%, 19%, 33.5%) are smaller than the corresponding GTR-JC69 cases for 17%, 22%, 42.5%. This is not surprising because the F81 model has more flexibility than JC69. Thus when the tree is true under the null hypothesis, which is actually to test model, the rejection rate is smaller.

When the generating tree is easy, Table 3.2 are the results of cases that either light or heavy among site variation and the null model is fixed as GTR. For both cases, the rejection rates under H_0^2 and H_0^3 easily reach 100%. Under H_0^1, where the tree is true, the rejection rates of light cases are 11%, 17.5% 27.5% and they are smaller than heavy case. This is because the model of light case is closer to the GTR than the heavy case. When the sequence lengths are 10000, The rejection rates reach 100% under H_0^1 for both cases.

Sequence Lengths	Hypothesis	Scenarios	
		GTR+D(heavy)-GTR	GTR+D(light)-GTR
300	H_0^1	27%	11%
	H_0^2	99%	99%
	H_0^3	98.5%	99.5%
500	H_0^1	28.5%	17.5%
	H_0^2	99%	100%
	H_0^3	100%	100%
1000	H_0^1	32.5%	27.5%
	H_0^2	100%	100%
	H_0^3	100%	100%
10000	H_0^1	100%	100%
	H_0^2	100%	100%
	H_0^3	100%	100%

Table 3.2: Rejection rates of each hypothesis in 2 scenarios based on easy simulation tree

3.3.3.2 The size and power of the test when the true tree is hard

Table 3.3 includes the results of 4 different scenarios: GTR-GTR, GTR-JC69, GTR-F81, GTR-HKY. For GTR-GTR, when the sequence lengths are less than 10000, the rejection rates under H_0^2 and H_0^3 are lower than the case where the true tree is easy tree. But they increase significantly as the sequence length becomes longer. The reason is when the generating tree is hard, it can impact the MLE of parameters and thus the expected probability of site patterns for the short sequence data. The sizes are all around 5% (6%, 5.5%, 5.5% and 4.5%) for any sequence lengths. The results demonstrate that when the true tree is hard, the power is lacking to reject the wrong tree topologies for the data with a relatively short sequence length.

The model misspecification scenarios include GTR-JC69, GTR-F81 and GTR-HKY. In the GTR-JC69 and the GTR-F81 cases where the LBA exists, the rejection rates under H_0^2, where the null tree is a LBA tree, are lower than those under the H_0^1 and the H_0^3. Because when the generating tree is hard and the null model is wrong, the estimated tree converges to be LBA tree. Hence, the H_0^2 is harder to reject. As the sequence length increases, the rejection rates under H_0^2 rise significantly. In the GTR-JC69 case, the rejection rate under H_0^2 is larger than that in the GTR-F81 since the degree of model misspecification is higher.

In the GTR-HKY, the degree of model misspecification is the smallest because the

GTR and HKY models are two similar substitution models. When the sequence length is less than 10000, the rejection rates of H_0^1, H_0^2 and H_0^3 are very close to those in the GTR-GTR. When the sequence length increases to 10000, the rejection rates under each hypothesis significantly increase (20%, 84%, 81%). Thus, it will approach 100% if the sequence length is long enough.

		Scenarios			
Sequence Lengths	Hypothesis	GTR-GTR	GTR-JC69	GTR-F81	GTR-HKY
300	H_0^1	6%	25.5%	17.5%	6%
	H_0^2	9%	21%	16%	9%
	H_0^3	9.5%	35.5%	24%	9%
500	H_0^1	5.5%	37.5%	29%	5.5%
	H_0^2	11%	26%	22%	10.5%
	H_0^3	12.5%	47%	32.5%	14%
1000	H_0^1	5.5%	67%	46%	7%
	H_0^2	15%	46%	39.5%	15.5%
	H_0^3	16%	78.5%	65%	16%
10000	H_0^1	4.5%	100%	100%	20%
	H_0^2	100%	100%	100%	84%
	H_0^3	100%	100%	100%	81%

Table 3.3: Rejection rates for each hypothesis in 4 analysis model scenarios when the true tree is hard tree

The results of the misspecification of heavy and light d are different (Table 3.4). Under the heavy case, when the null model is wrong and generating tree is hard, the LBA exists. The rejection rates under H_0^2 (null tree is the LBA tree) are lower than the H_0^1 and the H_0^3. In the light case, where the LBA is absent, the rejection rates are smaller than those in the heavy case. This indicates that the GTR model with light case of among site variation has less degree of model misspecification. The rejection rates eventually reach 100% when the sequence length increases to 10000 for both cases.

For some cases with model misspecification under hard tree, the rejection rates presented above give the power of this test as average over simulated data sets. Note that the number of the hypotheses which are rejected differ among individual data. Thus, this test may not provide an informative result when only one of the hypothesis (H_0^1, H_0^2, H_0^3) is rejeted. Hence, it is helpful to investigate how often all three (or only one) hypotheses are rejected within the simulation study. Now, I denote the rejection rate that only one hypothesis is rejected as R_1 and thus the R_2 and R_3 represent the rejection rate

Sequence Length	Hypothesis	Scenarios	
		GTR+D(heavy)-GTR	GTR+D(light)-GTR
300	H_0^1	63%	32%
	H_0^2	62%	37%
	H_0^3	68.5%	27.5%
500	H_0^1	77%	32.5%
	H_0^2	73.5%	43.5%
	H_0^3	82%	41%
1000	H_0^1	90%	29.5%
	H_0^2	82%	55%
	H_0^3	94%	41.5%
10000	H_0^1	100%	100%
	H_0^2	100%	100%
	H_0^3	100%	100%

Table 3.4: Rejection rates for each hypothesis in 2 analysis model scenarios when the true tree is hard tree

that two and three hypotheses are rejected. Some of the scenarios in the simulation study are used to present the results.

In GTR-F81 case, the rejection rates are R_1 = 9%, R_2=11% and R_3=10% when the sequence length is 300. When the sequence length increases to 1000, the rejection rates are R_1 = 13%, R_2=23.5% and R_3=27%. In GTR-JC69 case, the rejection rates are R_1 = 11%, R_2=15% and R_3=13% when the sequence length is 300 whereas when the sequence length increases to 1000, the rejection rates are R_1 = 11%, R_2=24% and R_3=38.5%. In the GTR+D(heavy)-GTR case, the rejection rates are R_1 = 2%, R_2=5% and R_3=46.5% when the sequence length is 300. When the sequence length increases to 1000, the rejection rates are R_1 = 3%, R_2=6.5% and R_3=78%. Hence, for the GTR+D(heavy)-GTR case, most of the cases the test could either reject all three trees or none of them. This is good, because rejecting only two of the trees could be misleading, suggesting that the other tree is the true tree.

3.3.4 Discussion

The goodness-of-fit test can reject combination of a wrong tree and a null model if the null model is the generating model with reasonable power for long sequences. The sizes of the test are at the nominal level for all different sequence lengths. When the null model is also wrong, the goodness-of-fit test has higher power to reject all combinations of wrong trees and wrong substitution models.

Another issue is that the test has to be performed for each possible tree , This requirement is difficult to meet if the number of taxa is large, then the number of possible tree topologies increase rapidly. A possible approach is to select a subset of the most possible tree topologies and cluster the site patterns according to the parsimony scores of each site pattern for this subset of tree topologies. I will not discuss the details for large taxa cases in this thesis.

In next section, another binning method will be developed which is unrestricted by tree topologies.

3.4 A goodness-of-fit test for substitution models

In this section, another way of binning site patterns will be developed so that Pearson's χ^2 test can be used for testing the substitution models. The null hypothesis is now:

H_0: The substitution model *M* is the true model.

The site patterns of the characters can be numerically summarized and cluster the site patterns using these numbers. This provides a better way for clustering analysis. One important factor in the substitution model is the equilibrium frequency of the nucleotide characters, $\pi = (\pi_A, \pi_C, \pi_G, \pi_T)$. The clustering of sites is based on the frequency summary statistics for each site. Thus for example, sites, (A, C, C, A), (C, A, C, A), (A, C, A, C) etc. will have the same summary statistics, and will be clustered together. A 4 taxa tree is used to demonstrate the method by binning the sites with equal frequency statistics together. The results of the goodness-of-fit tests will be presented using few simple scenarios and I will compare the size and power of the tests with LRT. This method is the developed to the large number of taxa cases. Finally, I will employ this method to analyze several real data sets.

3.4.1 Equal frequency binning for a 4 taxa tree

For a 4 taxa tree, there are 256 different site patterns. These site patterns are first classified into 5 different types according to the proportions of nucleotide characters. For example, sites XXYY and XYYX are the same type, because the proportions of X and Y are both $\frac{1}{2}$, and sites XXXY, XXYX are the same type because the proportions of X, Y are $\frac{3}{4}$ and $\frac{1}{4}$ respectively. Thus, there are 5 types of site patterns in total.

- Type 1: XXXX;
- Type 2: XXYY, XYYX, XYXY;
- Type 3: XXXY, XYXX, XXYX, YXXX;
- Type 4: XXYZ, XYZX, YZXX, XYXZ, ZXYX, ZXXY;
- Type 5: XYZW.

Each type of site patterns contains different number of bins depending on which nucleotide characters occupying X, Y, Z, and W. Thus, for each type, the number of bins are calculated as:
Type 1: $X \in \{A, C, G, T\}$; $\binom{4}{1} = 4$;
Type 2: $X, Y \in \{A, C, G, T\}, X \neq Y$; X, Y symmetric in the patterns; $\binom{4}{2} = 6$;
Type 3: $X, Y \in \{A, C, G, T\}, X \neq Y$; $\binom{4}{2} \times 2 = 12$;
Type 4: $X, Y, Z \in \{A, C, G, T\}, X \neq Y \neq Z$; Y, Z symmetric in the patterns; $\binom{4}{1}\binom{3}{2} = 12$;
Type 5: $X, Y, Z, W \in \{A, C, G, T\}, X \neq Y \neq Z \neq W$; $\binom{4}{4} = 1$;

Thus, there are 35 bins in total. With binning described above, the goodness-of-fit test procedure follows:

1. Calculate O_i as the observed count of the *i*th bin.
2. Compute the ML tree and MLE of parameters under the null model.
3. Calculate the expected probabilities of site patterns based on ML tree and MLE of model parameters.
4. Calculate expected probabilities, P_i, for each bin.
5. Calculate $E_i = nP_i$, where *n* is the sequence length.
6. The test statistic:
$$X_o^2 = \sum_{i=1}^{35} \frac{(O_i - nP_i)^2}{nP_i}$$
is compared to a χ^2 distribution with $df = 34$.

3.4.1.1 Simulation Design

200 data sets are simulated with sequence length 500 under the GTR, EF and EF+D models respectively. The parameters of GTR model are: $\pi_T = 0.308$, $\pi_C = 0.185$, $\pi_A = 0.308$, $\pi_G = 0.199$; $r_1 = 0.987, r_2 = 0.11, r_3 = 0.218, r_4 = 0.243, r_5 = 0.395, r_6 = 1$. The exchangeabilities of the EF model are the same as that of GTR model but equal frequencies $\pi_T = \pi_C = \pi_A = \pi_G = 0.25$ are assumed. The ratio of the branch lengths for generating each half of single sequence for the EF+D is 1:10. *INDELible1.03* is used for simulation.

In this simulation study, only the easy tree (shown in Fig 2.2 (a)) is used for simulation. Thus, the ML trees are the true tree. The simulation analysis scenarios in this section contain the EF-EF, GTR-EF, EF-JC69, GTR-JC69, and the EF+D-EF. The power of the test could vary depending on the degree of model misspecification. The likelihood ratio test (LRT) can only be used for comparisons of two nested models, thus I will use a large (e.g., GTR) alternative model for LRT. LRT is known to be the most efficient. As a reference, I will compare the power of this test to the LRT.

3.4.1.2 The size and power of the test

Table 3.5 lists the result of the goodness-of-fit test for each scenario. In the EF-EF, the goodness of fit test has 5.5% rejection rate, thus the size of the test is satisfactory. For the other cases with misspecification of the models (GTR-EF, EF-JC69, GTR-JC69), the rejection rates are all approximately 100%. Hence, the power is also satisfactory. For the case EF+D-EF, the rejection rate is 31%, the power of this test under this case is not as high as the other cases with model misspecification. Thus, this test is not very sensitive to the misspecification of the ASRV.

Scenarios				
EF-EF	GTR-EF	EF-JC69	GTR-JC69	EF+D-EF
5.5%	98%	100%	100%	31%

Table 3.5: The rejection rates of goodness-of-fit test for 4 scenarios

The results of LRT are used as reference and they are demonstrated in Table 3.6. The null models are EF and JC69. Since the EF and JC69 are both nested within the GTR model, the alternative models for LRT are chosen as the GTR model. The size of this test is 4.5%, which is slightly smaller than this test. For other cases, the rejection rates

are all 100%. When the true model is EF+D and EF is the null model, the rejection rate is 30.5% with the true model chosen as the alternative model.

Scenarios					
Generating Model	EF	GTR	EF	GTR	EF+D
Null Model	EF	EF	JC69	JC69	EF
Alternative Model	GTR	GTR	GTR	GTR	EF+D
	4.5%	100%	100%	100%	30.5%

Table 3.6: The rejection rates of LRT for the null models in 4 scenarios

In this simulation study, the scenarios contain different degrees of model misspecification. Size and power under the goodness-of-fit test are satisfactory for most cases in the simulation studies and they are similar to the LRT. Thus, this test seems to be a good tool for testing the adequacy of the model. One issue is that even though the binning method is simple with 4 taxa data, when the number of taxa increases, the binning method becomes difficult to apply because the number of site patterns increases rapidly. To deal with this issue, another binning procedure based on the same idea for the sequence with a larger number of taxa is developed.

3.4.2 Frequency based binning model test for large number of taxa

When the number of taxa *m* is large, there are, in theory, 4^m different site patterns. Binning based on exact equal frequency vectors is not practical for large *m* values. The idea is then extended such that sites with similar frequency vectors will be binned together. The *K*-means clustering method is used due to its simplicity.

In data mining, *K*-means clustering is a simple approach for clustering the observed (vector valued) data into different clusters according to their similarity, often measured by the Euclidian distance. Since the site patterns are all summarized by numerical values, it is easily to cluster these frequency vectors using any standard clustering method.

3.4.2.1 Binning Procedures

1. Summarize each site pattern into frequency vector $f_i = (f_{Ai}, f_{Ci}, f_{Gi}, f_{Ti}), i = 1, 2, \cdots, n$ and create an $n \times 4$ matrix:

$$F = \begin{pmatrix} f_{A1} & f_{C1} & f_{G1} & f_{T1} \\ f_{A2} & f_{C2} & f_{G2} & f_{T2} \\ \vdots & \vdots & \vdots & \vdots \\ f_{An} & f_{Cn} & f_{Gn} & f_{Tn} \end{pmatrix}$$

where each row contains the frequencies of observed nucleotides for the corresponding site.

2. The *K*-means clustering approach is used for binning the rows in matrix *F* into *K* bins.

3. For $j = 1, 2, \cdots, K$, denote the center of *j*th bin as C_j. Calculate the observed frequency for *j*th bin, O_j, as the counts of all sites assigned to *j*th bin.

4. Compute the ML tree and the MLE for all parameters.

5. Parametric bootstrap is used to simulate an extremely long (*M* sites) DNA sequence data X^* based on the ML tree and the MLE of model parameters.

6. From sequence data X^*, calculate the $M \times 4$ frequency matrix F^*, where each row contains the frequencies of nucleotide characters of each site:

$$F^* = \begin{pmatrix} f^*_{A1} & f^*_{C1} & f^*_{G1} & f^*_{T1} \\ f^*_{A2} & f^*_{C2} & f^*_{G2} & f^*_{T2} \\ \vdots & \vdots & \vdots & \vdots \\ f^*_{AM} & f^*_{CM} & f^*_{GM} & f^*_{TM} \end{pmatrix}$$

7. Cluster the rows in F^* to *K* clusters by comparing the Euclidian distance of each row to the *K* centers calculated in step 3 $(C_1, C_2, \cdots, C_K)$ and assigning the row to the cluster with the smallest Euclidian distance. Denote the number of rows assigned to the *j*th bin as S_j. Then, the expected size of the *j*th bin E_j, can be calculated as:

$$E_j = \frac{nS_j}{M}$$

where *n* is the sequence length in the observed data set.

8. The test statistic is:

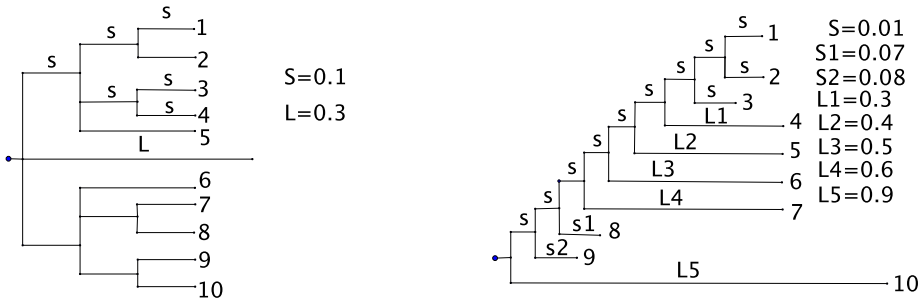
$$X_o^2 = \sum_{j=1}^{n} \frac{(O_j - E_j)^2}{E_j}$$

Under H_0, X_o^2 follows the χ^2 distribution with $df = K - 1$

3.4.2.2 Simulation design

In this simulation, two 10 taxa tree topologies, which are referred to as symmetric and asymmetric trees respectively, are used for generating DNA sequences. Figure 3.3 (a) is a symmetric 10 taxa tree and (b) is an asymmetric 10 taxa tree with specified branch lengths. A symmetric tree can be called an easy estimation problem whereas the asymmetric tree is called a "harder estimation problem", where the ratio of correct estimation tree to incorrect estimation tree is 3:1. For each of the two tree topologies, I employed models GTR, F81, and GTR+D to simulate data. 200 data sets with sequence length fixed at 500 are simulated for each scenario. *INDELible1.03* is used for simulation. The following generating and analysis model pairs are used to find both the size and power of the test.

There are 6 different scenarios: (1) GTR-GTR (No model misspecification), (2) GTR-F81, (3) F81-JC69, (4) GTR-HKY, (5) GTR+D(heavy)-GTR (6) GTR-JC69. The results in terms of the rejection rates at 5% significance level among the 200 data are presented below.



(a) Symmetric Tree (b) Asymmetric Tree

Figure 3.3: 10 taxa simulation trees

3.4.2.3 Analysis results

In the cluster analysis of the data, an optimal *K* can be decided by some algorithm and data together. However, this may result in different *K* values for different data sets in the

simulation. For simplicity, I fix K to 7, 30, and 70 in the following analysis. It can be observed from the analysis results that the test is not very sensitive to the K values in most cases.

3.4.2.4 Results for symmetric tree

The results of GTR-GTR case demonstrate that the size of the test are 0.5%, 0%, 0% when K is 7, 30 and 70 (Table 3.7). For the GTR-JC69 case, the model misspecification is the highest, the rejection rates are all 100% for any value of K. Under the GTR-F81, the rejection rates are 100%, 99% and 100% when K is 7, 30, and 70 respectively. There are similar results for the F81-JC69 case, which has the rejection rates 100%, 99.5% and 98% when K=7, 30, and 70, respectively. Under GTR-HKY, the rejection rates are 20%, 43% and 34%, which are not very high since the GTR-HKY has the smallest degree of model misspecification. But when K value is selected to be 30 and 70, the rejection rates are significantly higher than for K = 7. For the GTR+D - GTR case, the rejection rates are 20%, 69% and 93%, which also increase as K becomes larger. In most simulation cases, this test has enough power to reject the wrong models.

	Scenarios					
	GTR-GTR	GTR-JC	GTR-HKY	GTR-F81	F81-JC	GTR+D-GTR
K=7	0.5%	100%	20%	100%	100%	20%
K=30	0%	100%	43%	99%	99.5%	69%
K=70	0%	100%	34%	100%	98%	93%

Table 3.7: Rejection rates of each hypothesis in 6 scenarios based on 10 taxa symmetric tree for three K values

3.4.2.5 Results for asymmetric tree

For simulations based on the asymmetric tree, the results are similar to the results under the symmetric tree (Table 3.8). In the GTR-GTR case, the rejection rates are 0%, 0% and 1% when K = 7, 30 and 70. Thus, the sizes of the test are all good regardless of the value of K. In case of misspecification of the substitution model, the test has enough power to reject the wrong models. In the F81-JC69 case, the rejection rates are 100%, 100% and 99.5% and 100%, 99% and 100% for the GTR-F81 case, when K = 7, 30 and 70. In the GTR-JC69 case, the rejection rates are always 100% regardless of the K values. In the GTR+D-GTR case, the rejection rates depend on the K values. When K =

7, the rejection rate is 19% and it increases to 93% and 100% when K = 30 and K=70 respectively. For the GTR-HKY, the rejection rates demonstrate that HKY model is hard to reject. The rejection rates are 10.5%, 45% and 52.5% for K =7, 30 and 70 respectively. Again, the rejection rates in GTR-HKY case depend on the K values, but the power is lower comparing to other model misspecification cases. Hence, for the DNA sequences generated under asymmetric tree, the results are similar to that of symmetric tree.

	Scenarios					
	GTR-GTR	GTR-JC	GTR-HKY	GTR-F81	F81-JC	GTR+D-GTR
K=7	0%	100%	10.5%	100%	100%	19%
K=30	0%	100%	45%	99%	100%	93%
K=70	1%	100%	52.5%	100%	99.5%	100%

Table 3.8: Rejection rates in 6 scenarios based on 10 taxa asymmetric tree for three K values

In summary, the proposed test with binning based on frequencies has good power for both the symmetric and the asymmetric trees when the substitution model is misspecified, and it can be used for a tree with larger number of taxa. For some cases, an appropriate K value should be determined in order to draw the correct conclusion.

3.4.2.6 Discussion

Based on the simulation study, the goodness-of-fit test has satisfactory size and enough power to reject the wrong models in different scenarios. The power of the goodness-of-fit test based on K-means clustering on the site frequency vectors is promising. The procedure of binning the site patterns is simple and thus can be easily applied. Comparing with LRT based model adequacy tests, the model assumptions are simpler because the LRT test requires both appropriate null and alternative models. Jeniffer and Sullivan (2010) have examined the GC test for many empirical data. I will employ the goodness-of-fit test for 3 empirical data in the next section.

3.5 Empirical data analysis

3.5.1 Data collection and the hypothesis test

The three empirical data sets are selected among 25 empirical data in Ripplinger and Sullivan (2010). The matrix ID of these data sets are M1000, M780 and M2309. Jeniffer

and Sullivan (2010) and Goldman (1993) pointed out that the GC test failed to reject the JC+$I + \Gamma$ for many empirical data, where "I" represents the proportion of invariant sites, and the Γ represents the among site variation rate. Here, I will apply this test on the same type of models which incorporate with $I + \Gamma$. The null hypotheses for each of the data sets are:

$H_0 1$: The JC+$I + \Gamma$ is the true model;
$H_0 2$: The F81+$I + \Gamma$ is the true model;
$H_0 3$: The HKY+$I + \Gamma$ is the true model;
$H_0 4$: The GTR+$I + \Gamma$ is the true model;

For each data, the optimal *K* can be different. Peeples (2011) reviewed the solutions for choosing optimal *K* in clustering analysis. Here, I will use the most common solution as follows: to choose an optimal *K*, I compare the Sum of Squared Errors (SSE) for different *K* values. The SSE for each cluster is defined as the sum of squared distances between each element of a cluster with the centre of this cluster. Thus, as a global measurement of errors, SSE is the sum of SSE's over all clusters. As *K* increases, the SSE will decrease since the sizes of the clusters are smaller. Then I can create a plot of SSE against sequential *K* values. The optimal *K* is the elbow point at which the reduction of SSE becomes slow dramatically.

In this analysis, I will examine the p-values of the test under the optimal *K* as well as some other *K* values which are close to the optimal *K*.

A DNA sequence data set with 100000 sites are simulated based on parametric bootstrap procedure under each hypothesis. The p-values of the GC test for each data have been provided in the supplementary material of the Jennifer and Sullivan (2010) and are recorded here to compared with the p-values of this test.

3.5.2 Results of empirical data analysis

3.5.2.1 DATA 1: M1000

The M1000 (Cox, Huynh, and Stone 1995) consists of 10 taxa and the sequence length is 817. The plot of SSE against *K* is shown in Fig 3.4: The optimal *K* is 4 and I will also include *K*=5,6,7,8,9. The p-values for *K* equals to 4 to 9 are listed in Table 3.9. Under $H_0 1$, all of the p-values are less than 0.01 regardless of *K* values. Under $H_0 2$, only when

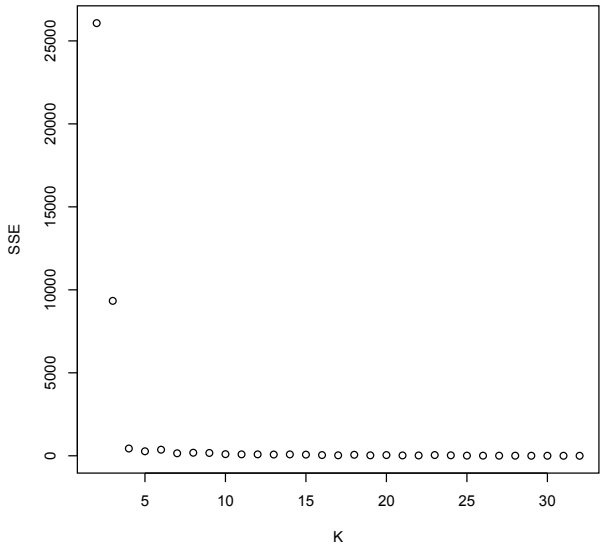


Figure 3.4: Plot of SSE against K for M1000

		Hypothesis			
		$H_0 1$	$H_0 2$	$H_0 3$	$H_0 4$
GC Test		0.01	0.04	0.24	0.1
Goodness-of-Fit Test	K=4	p<0.01	0.76	0.77	0.84
	K=5	p<0.01	0.82	0.85	0.95
	K=6	p<0.01	0.91	0.87	0.96
	K=7	p<0.01	0.79	0.82	0.89
	K=8	p<0.01	0.67	0.79	0.87
	K=9	p<0.01	p<0.01	0.86	0.88

Table 3.9: p-values of GC and goodness-of-fit test under each hypothesis for M1000

K=9, the p-values is less than 0.01. Thus, I can still draw conclusion that $H_0 2$ is rejected. Under $H_0 3$ and $H_2 4$, the p-values are large regardless of *K*. The p-values of the GC test (Table 3.9) show that $H_0 1$ and $H_0 2$ can also be rejected. Thus, the goodness-of-fit test and GC test can reach the same conclusion for this data.

3.5.2.2 DATA 2: M780

The M780 (Leander and Porter 2000) consists of 10 taxa and the sequence length is 199. The plot of SSE against *K* is shown in Figure 3.5. The optimal *K* is 6. The p-values

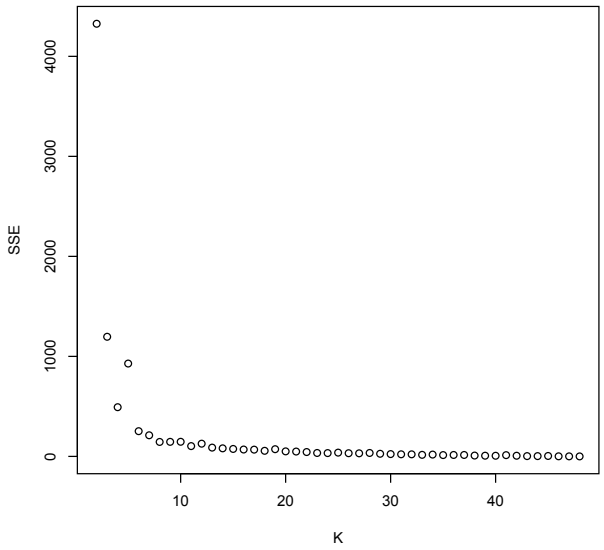


Figure 3.5: Plot of SSE against *K* for M780

under $H_0 1$ are all smaller than 0.01. But for other hypotheses, the p-values are all greater than 0.1, thus I can only reject the $H_0 1$ for this data.

The p-values of the GC test (Table 3.10) show that all of p-values are greater than 0.1, thus GC test fails to reject the $H_0 1$.

3.5.2.3 DATA 3: M2309

The M2309 (Brandli, Handley, Vogel and Perrin 2005) consists of 11 taxa and the sequence length is 374. The plot of SSE against *K* is shown in Figure 3.6.

The optimal *K* is 7, I also examine *K*=8, 9, 10 in the results. Under $H_0 1$, all of the p-values are less than 0.01. Under $H_0 2$, only when *K*=10, the p-value is small enough to reject the hypothesis. Under $H_0 3$, when *K*=8, the p-value is 0.06, where the

		Hypothesis			
		$H_0 1$	$H_0 2$	$H_0 3$	$H_0 4$
GC Test		0.84	0.85	0.85	0.85
Goodness-of-Fit Test	K=6	p<0.01	0.35	0.57	0.43
	K=7	p<0.01	0.32	0.47	0.39
	K=8	p<0.01	0.24	0.33	0.31

Table 3.10: p-values of GC test and goodness-of-fit test under each hypothesis for M780

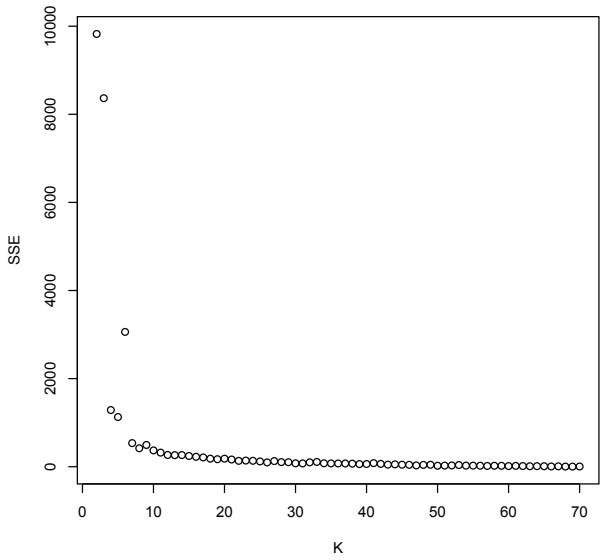


Figure 3.6: Plot of SSE against K for M2309

hypothesis can be marginally rejected. Thus, by adjusting *K*, I can reject $H_0 1$, $H_0 2$ and $H_0 3$ marginally.

The p-values of the GC test are all less than 0.01 except $H_0 4$. Hence, the results of GC test are similar to the goodness-of-fit test.

		Hypothesis			
		$H_0 1$	$H_0 2$	$H_0 3$	$H_0 4$
GC Test		p<0.01	p<0.01	p<0.01	0.16
Goodness-of-Fit Test	*K*=7	p<0.01	0.71	0.32	0.89
	K=8	p<0.01	0.51	0.06	0.25
	K=9	p<0.01	0.12	0.14	0.14
	K=10	p<0.01	0.04	0.18	0.36

Table 3.11: p-values of GC and goodness-of-fit test under each hypothesis for M2309

3.6 Conclusion

In this chapter, I explored the performance of two binning methods for Pearson's goodness-of-fit test. The hypothesis based on the first method consists of the assumption of tree and substitution model. Under the easy 4 taxa tree, this method presents the power to reject the wrong tree when the null model is correct. For any case with model misspecification, the power is not satisfied when the sequence length is insufficient. For the data generated under the hard tree case, the power is generally less than those in easy tree. For any cases with model misspecification, the rejection rates under each tree reach 100% only when the sequence length is 10000. In general, this approach has limited power. If the number of taxa is large, it is difficult to use this binning method.

Based on the second binning method, the goodness-of-fit test can be applied for sequence with larger number of taxa to test the substitution model. The results of this test show that the power significantly increases for most cases in simulation. For real data analysis, the results of analysis of M1000 is consistent with the GC test. For the M780, the test has more power than the GC test because GC test fails to reject the JC+$I + \Gamma$ model which can be rejected by this test. For M2309, this test can marginally reject the HKY+$I + \Gamma$ model and it is close to the conclusion drawn by GC test. However, the this test require the MLE under the ML tree, hence, it may also sensitive to the tree (when the ML the tree is extremely wrong, it can impact on the results).

CHAPTER 4

CONCLUSION AND FUTURE WORKS

4.1 Conclusion

The ML method provides consistent results for the data generated from an easy tree, regardless of the analytical model. For the data generated under a hard tree, the results can also converge to the true tree if the analytical model is correct. However, for the data generated under a hard tree and analyzed with misspecified models, the results generally converge to the LBA tree except in some special cases when models are only slightly misspecified. As a reference of the ML method, the EL method sometimes can better demonstrate these results. Within the LBA scenarios, even when the ML tree is the true tree, the branch length estimates are often very biased.

I developed two methods based on Pearson's goodness-of-fit test for testing the adequacy of substitution models. The basic ideas for these two tests are both to make Pearson's goodness-of-fit test applicable through binning the site patterns. Two different binning methods have been developed in this thesis: (1) binning depending on the tree in the null hypothesis and (2) binning based on the frequencies of the nucleotide characters of each site. The first test has acceptable size but limited power for shorter sequence lengths.

Based on the second binning method, the test can be easily applied to the data with large number of taxa. This test has both satisfactory sizes and powers in most simulation scenarios for the sequences with either low or high number of taxa. In the empirical data analysis, this test was compared to the GC test, and the results show that it has similar or larger power to reject the null models.

4.2 Future work

The second test has shown to be a simple and powerful model adequacy test. The K value in this test is directly related to the power of the test. In future work, better methods to choose the optimal K value to maximize the power of the test should be developed. Some explanations about the high power of this test should be explored so that this test can be better accepted and applied. The sensitivity of the test to the tree topology should also be investigated. The design of the test should make it relatively robust to misspecification of topology. However this should be confirmed by simulation studies.

BIBLIOGRAPHY

[1] J. Bergsten. 2005. "A review of long-branch attraction." *Cladistics* 21:163-193.

[2] L. Brandli, L. J. Handley, P. Vogel and N. Perrin. 2005. "Evolutionary history of the greater white-toothed shrew (Crocidura russula) inferred from analysis of mtDNA, Y and X chromosome markers." *Mol. Phylogenet. Evol.* 37:832-844.

[3] W. J. Bruno and A. L. Halpern. 1999. "Topological bias and inconsistency of maximum likelihood using wrong models." *Mol. Biol. Evol*, 16:564-566.

[4] L. L. Cavalli-Sforza and A. W. F. Edwards. 1967. "Phylogenetic analysis: Models and estimation procedures." *Evolution* 32:550-570.

[5] P. Cox, K. Huynh and B. Stone. 1995. "Evolution and systematics of the Pandanaceae." In: Rudall P., Cribb P., Cutler D., & Humphries C., eds. *Monocotyledons: systematics and evolution.* pp. 663-684. Kew, Royal Botanic Gardens.

[6] A. W. F. Edwards and L. L. Cavalli-Sforza. 1963. "The reconstruction of evolution."*Annals of Human Genetics* 27: 105-106 (also published in *Heredity* 18: 553).

[7] A. W. F. Edwards and L. L. Cavalli-Sforza. 1964. "Reconstruction of evolutionary trees." pp. 67-76 in *Phenetic and Phylogenetic Classification*, ed. V. H. Heywood and J. McNeill. Systematics Association Publ. No.6, London.

[8] J. Felsenstein. 1978. "Cases in which parsimony or compatibility methods will be positively misleading." *Syst. Zool*. 27:401410.

[9] J. Felsenstein. 1981. "Evolutionary trees frome DNA sequences: Maximum likelihood approach." *Journal of Molecular Evolution* 17: 368-376.

[10] J. Felsenstein. 2004. *Inferring Phylogeny*. Sinauer.

[11] W. Fletcher and Z. H. Yang. 2009. "INDELible: A flexible simulator of biological sequence evolution." *Mol. Biol. Evol* 26 (8): 1879-1888.

[12] N. Goldman. 1993. "Statistical tests of models of DNA substitution." *J. Mol. Evol.* 36:18298.

[13] M. Hasegawa, H. Kishino and T. Yano. 1985. "Dating of human-ape splitting by a molecular clock of mitochondrial DNA." *J. Mol. Evol* 22: 160-174.

[14] J. Huelsenbeck. 1995. "Performance of phylogenetic methods in simulation." *Syst. Biol.* 44(1): 17-48.

[15] T. H. Jukes and C. R. Cantor. 1969. "Evolution of protein molecules." pp. 21-132 in *Mammalian Protein Metabolism*, Vol. III, ed. M. N. Munro. Academic Press, New York.

[16] S. Kullback and R. A. Leibler. 1951. "On information and sufficiency." *Ann. Math. Statist.* 22, 79-86.

[17] C. Lanave, G. Preparata, C. Saccon and G. Serio. 1984. "A new method for calculating evolutionary substitution rates." *Journal. Mol. Evol* 20: 86-93.

[18] C. A. Leander and D. Porter. 2000. "The Labyrinthulomycota is comprised of three distinct lineages." *Mycologia* 93:459-464.

[19] M. A. Peeples. 2011. "R Script for K-Means Cluster Analysis." [online]. Available: *http://www.mattpeeples.net/kmeans.html*.

[20] D. Posada and K. A. Crandall. 1998. "ModelTest: testing the model of DNA substitution." *Bioinformatics* 14(9): 817-818.

[21] J. Ripplinger and J. Sullivan. 2010. "Assessment of Substitution Model Adequacy Using Frequentist and Bayesian Methods." *Mol. Biol. Evol.*, 27 (12):2790-2803.

[22] J. Sullivan and P. Joyce. 2005. "Model Selection In Phylogenetics." *Annu. Rev. Ecol. Evol. Syst.*, 2005. 36: 445 - 66.

[23] E. Susko. 2011. "Large sample approximations of probabilities of correct evolutionary tree estimation and biases of maximum likelihood estimation." *Statistical Applications in Genetics and Molecular Biology* 10(1), Article 10.

[24] P. J. Waddell, R. Ota and D. Penny. 2008. "Measuring Fit of Sequence Data to Phylogenetic Model: Gain of Power using Marginal Tests." *Mol. Biol. Evol.* 22: 395.

[25] Z. H. Yang. 1994. "Estimating the Pattern of Nucleotide Substitution." *J. Mol. Evol.* 39: 105-111.