A NEW TEST TO BUILD CONFIDENCE REGIONS USING BALANCED MINIMUM EVOLUTION

by

Wei Dai

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

at

Dalhousie University
Halifax, Nova Scotia
August 2013

Table of Contents

List of Tables

List of Figures

Acknowledgements

I would like to express my appreciation to Dr. Hong Gu and Dr. Toby Kenney for their supervision on this thesis. This thesis could not have been finished without their guidance and reviews.

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

Thanks to my mother and father, their love and care has helped me to get through many lonely evenings.

Abstract

In phylogenetic analysis, an important issue is to construct the confidence region for gene trees from DNA sequences. Usually estimation of the trees is the initial step. Maximum likelihood methods are widely applied but few tests are based on distance methods. In this thesis, we propose a new test based on balanced minimum evolution. We first examine the normality assumption of pairwise distance estimates under various model misspecifications and also examine their variances, MSEs and squared biases. Then we compare the BME method with the WLS method in true tree reconstruction under different variance structures and model pairs. Finally, we develop a new test for finding a confidence region for the tree based on the BME method and demonstrate its effectiveness through simulation.

Chapter 1

Introduction

Phylogeny is the science of studying evolutionary history of different organisms. Cavalli-Sforza and Edwards (1967) indicated that the phylogeny problem was actually a statistical inference problem. A phylogenetic tree is a tree diagram representing phylogenetic relationships among a group of organisms. Different tree reconstruction methods have been developed to estimate phylogenetic trees from genetic sequence data.

Three different types of approaches to phylogeny reconstruction have been widely used so far, namely parsimony methods, maximum likelihood (ML) methods and distance-based methods. Each has its own strengths and weaknesses. Parsimony methods work in the following way: given genetic sequence data, there are many phylogenetic trees available and parsimony methods choose the most parsimonious one, i.e the one such that fewest evolutionary changes are needed to generate the data. Parsimony is a simple approach, however, sometimes it is not statistically consistent such as in the long branch attraction case (Felsenstein 1978). That means it is not guaranteed to output the correct tree even if sufficient sequence data are given. ML methods use a stochastic model of sequence evolution that describes the probabilities of substitutions. Given the substitution model, the ML tree is the tree that maximizes the probability of the sequence data. This probability is called the likelihood of the data. ML is consistent (Wald 1949; Felsenstein 1973; Yang 1994) and more powerful than parsimony because it employs an explicit model for character evolution. The drawback of ML methods is that it is impractical when dealing with large data sets, because the tree space, i.e. the collection of all possible trees, grows exponentially with the number of sequences (Felsenstein 2004). Distance-based methods were introduced by Cavalli-Sforza and Edwards (1967) and are the only known methods that can handle the data from thousands of sequences. The general idea is to calculate a measure of distance between each pair of species, and then

form a tree that can best approximate the calculated pairwise distances. Distance-based methods are mostly consistent and require to search the whole tree space too. There are a varieties of distance-based methods developed, for example UPGMA and neighbor-joining (Saitou and Nei 1987) for computation speed, and weighted least-squares (Fitch and Margoliash 1967) for accuracy. The balanced minimum evolution (BME) is another distance-based method that is at least as fast as neighbour-joining, and as accurate as weighted least-squares (Desper and Gascuel 2002). BME is the focus of this thesis.

As the basis of distance-based methods, the accuracy of pairwise distance estimates is fundamental. Most often, the pairwise distance estimates are also based on the nucleotide substitution models. There are many nucleotide substitution models developed so far; see Felsenstein (2004) for a survey of the most commonly used models. But due to statistical uncertainty and limitations of finite data, the pairwise distance estimates may not be accurate enough and the phylogeny reconstruction may not always reveal the true evolutionary history.

After the introduction in this chapter, we will first examine the accuracy of pairwise distance estimates with model misspecification through simulations. Then we will compare the performances of weighted least-squares (WLS) and balanced minimum evolution (BME) in tree estimation. Finally, we propose a new method to build the confidence region for phylogenetic trees based on BME and test its effectiveness.

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 . The total number of possible m taxa rooted tree topologies is:

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

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

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

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 consists of 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 four taxa a, b, c and d. The first 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 is essentially a series of changes of nucleotides in the DNA sequences of their ancestors. There are many substitution models to describe the changes in DNA sequences.

1.2 Substitution 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 possible changes in total. A change is called a *transition* if it occurs within either pyrimidine (cytosine (C) and thymine (T)) or purine (adenine

(A) and guanine (G)) 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:

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

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

Each row of above matrix sums to 1.

1.2.2 Substitution rate matrix

In distance methods the pairwise distances are sometimes calculated by maximum likelihood (ML). Estimating pairwise distance requires a probability substitution matrix, which can be determined by a rate matrix, denoted as Q (Table 1.3). Each element of Q is considered as rate of exchange at an instant time dt between nucleotide characters. The entries q_{ij} of matrix Q can be expressed as a product of equilibrium frequency of nucleotide j, π_j and exchangeability r_{ij}. Thus $q_{ij} = r_{ij}\pi_j$, for $i \neq j$; and $q_{ii} = -\sum_{i \neq j} q_{ij}$.

The matrix Q of a substitution model depends on the particular assumptions of elements in $R = (r_{ij})$ and $\pi = (\pi_C, \pi_A, \pi_G, \pi_T)$. Table 1.4 shows the parametrization of Q matrix when reversibility is assumed. Five different DNA models will be applied in this thesis: JC69, F84, HKY85, GTR and GTR + Γ, thus we review their rate matrices and estimates of pairwise distance under each model as follows.

	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	-	$r_1\pi_C$	$r_2\pi_A$	$r_4\pi_G$
C	$r_1\pi_T$	-	$r_3\pi_A$	$r_5\pi_G$
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: The GTR Q matrix used for modeling substitution

1.2.3 Substitution models

JC69 model

The JC69 (Jukes and Cantor, 1969) is the simplest substitution model in phylogeny because it assumes the exchangeabilities and character frequencies are all constant. Thus, the matrix Q is given by Table 1.5. The pairwise distance esitmate has a closed form, given as $-\frac{3}{4}\log(\frac{4}{3}(\hat{p}-\frac{1}{4}))$ where $\hat{p}$ is the proportion of sites with two nucleotides the same. (Yang, 2006)

	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

HKY85 model

The HKY85 (Hasegawa, Kishino and Yano, 1985) is an extension of JC69 model, where character frequencies of A, C, G, T, $(\pi_A, \pi_C, \pi_G, \pi_T)$, are not restricted and the Q matrix depends on both character frequencies and *transition-transversion* ratio κ. Hence, the Q matrix is as given in Table 1.6. In PAML (Yang, 1994), under HKY85

model, the estimate of pairwise distance is calculated according to the following formulae, which is a special case of TN93 model formulae (Yang, 2006).

$$2\pi_T\pi_C\kappa b + 2\pi_A\pi_G\kappa b + 2\pi_Y\pi_G b$$

where

$$\kappa = \frac{a_1 - \pi_R b}{\pi_Y b},$$

$$b = -\log(1 - \frac{V}{2\pi_Y\pi_R}),$$

$$a_1 = -\log(1 - \frac{\pi_Y S_1}{2\pi_T\pi_C} - \frac{V}{2\pi_Y})$$

V denotes the proportion of sites with transversional differences, S_1 denotes the proportion of sites occupied by two different pyramidines (i.e sites occupied by CT or TC) in the two sequences; $\pi_R = \pi_A + \pi_G$ and $\pi_Y = \pi_C + \pi_T$ as total frequencies of purines and pyrimidines respectively. (Yang, 2006).

	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 1.6: Q matrix of HKY85

F84 model

The F84 (Felsenstein 1984) model is similar to HKY85 model. It also assumes that the character frequencies are not restricted but the transition rate ratios are assumed to be different for purines and pyrimidines. The matrix Q has the form given in Table 1.7.

	T	C	A	G
T	-	$(1 + \kappa/\pi_Y)\pi_C$	π_A	π_G
C	$(1 + \kappa/\pi_Y)\pi_T$	-	π_A	π_G
A	π_T	π_C	-	$(1 + \kappa/\pi_R)\pi_G$
G	π_T	π_C	$(1 + \kappa/\pi_R)\pi_A$	-

Table 1.7: Q matrix of F84

The pairwise distance esitmate is calculated as

$$2(\frac{\pi_T\pi_C}{\pi_Y}+\frac{\pi_A\pi_G}{\pi_R})a-2(\frac{\pi_T\pi_C\pi_R}{\pi_Y}+\frac{\pi_A\pi_G\pi_Y}{\pi_R}-\pi_Y\pi_R)b$$

with

$$\hat{\kappa}=\frac{a}{b}-1,$$

where

$$a=-\log\left(1-\frac{S}{2[(\pi_T\pi_C/\pi_Y)+(\pi_A\pi_G/\pi_R)]}-\frac{[(\pi_T\pi_C\pi_R/\pi_Y)+(\pi_A\pi_G\pi_Y/\pi_R)]V}{2(\pi_T\pi_C\pi_R+\pi_A\pi_G\pi_Y)}\right),$$

$$b=-\log(1-\frac{V}{2\pi_Y\pi_R})$$

S denotes the proportion of sites with transitional differences, V denotes the proportion of sites with transversional differences (Yang, 2006).

GTR (Generalized time-reversible) model

The GTR model (Tavare 1986) uses the most general form of the rate matrices subject to the reversibility condition. The reversibility condition means the evolutionary process is the same when substution runs from past to the present or from present back to the past i.e $\pi_i q_{ij}=\pi_j q_{ji}$ (Yang 1994). The reversibility condition holds for all of the models we consider in this thesis. In the GTR model, both π and R are fully parametrized. The Q matrix of the GTR model is given in Table 1.4 with exchangeability parameters $(r_1,r_2,r_3,r_4,r_5,r_6)$ where $r_6=1$ is fixed.

GTR + Γ model

In addition to the above models describing the rates of change between nucleotides, some models have further assumptions for rate variation among sites in a sequence, i.e often we do not know the rate of each site in a sequence, but we assume a distribution of rates, and each site has a rate drawn from that distribution at random (Felsenstein 2004), take rates in table 1.4 as example, all r_i are random varaibles from a distribution. The gamma-distributed-rate model (Uzzell and Corbin 1971) is the most used continuous model. It allows the rate variation among sites to follow gamma distribution. The gamma density function is

$$f(r)=\frac{1}{\Gamma(\alpha)\beta^\alpha}r^{\alpha-1}e^{\frac{-r}{\beta}}$$

where the α and β are parameters. The mean of the distribution is $\alpha\beta$, and the variance is $\alpha\beta^2$. α is called a "shape parameter" here. If the mean is fixed, when α increases the variance decreases, and vice versa. Thus small α values can be used to model large rate variation among sites. The Γ distribution here is only used for the mathematical convenience. It has no significant biological meaning (Felsenstein 2004). If the rate variation among sites in the GTR model follows the Γ distribution then it is called GTR + Γ model. The Γ model used in practice always constrains $\beta = 1/\alpha$, so that the mean is 1. This allows edge-lengths to be interpreted as expected numbers of substitutions.

1.3 Summary of the thesis

This thesis examines the pairwise distance estimates in distance methods and compares the BME method with the WLS method in tree topology estimation. It further develops a new method to construct confidence regions based on the BME method. More specifically, Chapter 2 studies the pairwise distance estimates with model misspecification, investigates the bias, variance and mean square error of the pairwise distance estimates. Chapter 3 thoroughly compares the BME and WLS methods in estimating tree topology under different scenarios. Chapter 4 introduces a new method to calculate the confidence regions based on BME and tests its effectiveness. Chapter 5 concludes the thesis.

Chapter 2

The pairwise distance estimates

Distance-based methods require pairwise distances calculated between each pair of species. We rely on evolutionary models to compute pairwise distances, but if the model used to analyze the data is misspecified, especially when the true tree is a hard tree with large distances between taxa, the pairwise distances may not be accurately estimated. This may result in not finding the true tree topology (Olsen, 1987; Lockhart et al., 1994; Van de Peer et al., 1996; Susko et al., 2004; Sullivan and Joyce, 2005). Some distance-based methods, such as the ordinary least squares method, implicitly assume the pairwise distance estimates are independent and normally distributed (Nei 1996). The normal distribution assumption approximately holds for the pairwise distance estimate when the sequence length is large. For example, under the JC69 model, the pairwise distance estimate is based on the formula $-\frac{3}{4}\log(\frac{4}{3}(\hat{p}-\frac{1}{4}))$. When $\hat{p}$ is an MLE, then the $\hat{p}$ approximately follows normal distribution. Then according to delta method, the pairwise distance esitmate should also approximately follow the normal distribution. This is true even when JC69 model is a misspecified model.

In this chapter, we will examine the estimates of pairwise distances with and without model misspecifications through simulation studies. For four 5-taxon trees with different tree branch lengths, we examine normality assumptions through Q-Q plots of the estimated pairwise distances. We also check their mean squared errors (MSE), variances and biases against true pairwise distances. We investigate the relationship between the standard deviation (SD) and the mean of the estimates.

2.1 Simulation design

The purpose of simulation is to examine some aspects of pairwise distance estimates, namely (1) the normality assumption, (2) MSEs, biases and variances of distance estimates and how they vary against the true distance values, (3) how the SDs are

related to the means of distance estimates. To achieve these goals, we use four 5-taxon trees shown in Figures 2.1 and 2.2 as true tree topologies. The lengths of these trees are 3.6, 4.8, 3.2 and 5.9 for tree 1, tree 2, tree 3 and tree 4 respectively.

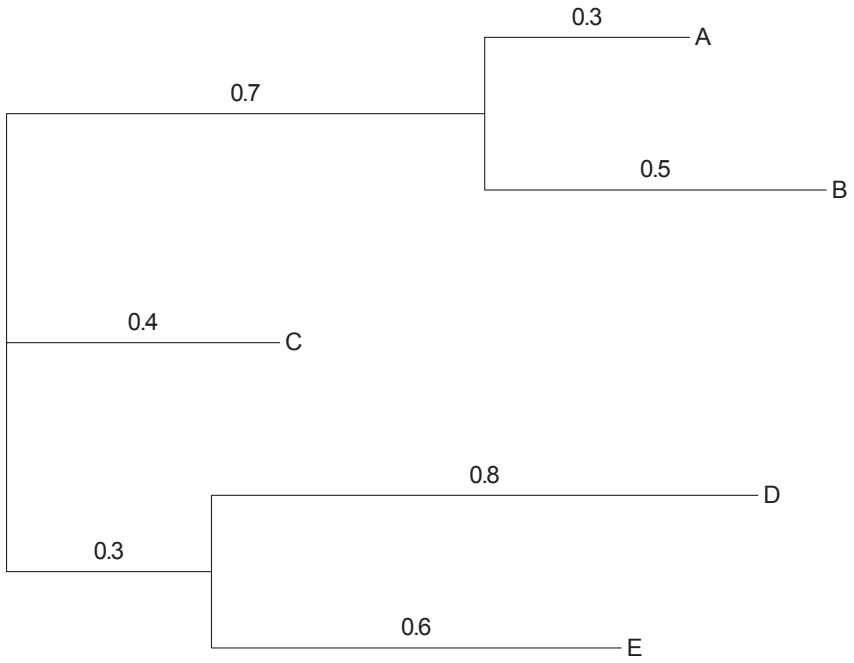
We simulate data using the trees in Figures 2.1 and 2.2 and models JC69, GTR and GTR+Γ; we analyze the data using models F84 and JC69 , so there are six different "simulation-analysis" model pairs in total, i.e. GTR-F84, GTR-JC69, GTR+Γ-F84, GTR+Γ-JC69, JC69-JC69, and JC69-F84. Except for JC69-JC69, all others are misspecified models. We use the Indelible1.03 (Fletcher and Yang 2009) to simulate data; the sequence length of each data set is 500, and 1000 replicates are simulated for each case. For each scenario we use DNADIST from the phylip package (Felsenstein 1980) to analyze these simulated sequence data. The parameters used for GTR models are $\pi_T = 0.1$, $\pi_C = 0.2$, $\pi_A = 0.3$, $\pi_G = 0.4$; $r_1 = 0.2, r_2 = 0.4, r_3 = 0.6, r_4 = 0.8, r_5 = 1.2$. The α value used for the GTR+Γ model is 0.385.

2.2 Simulation results

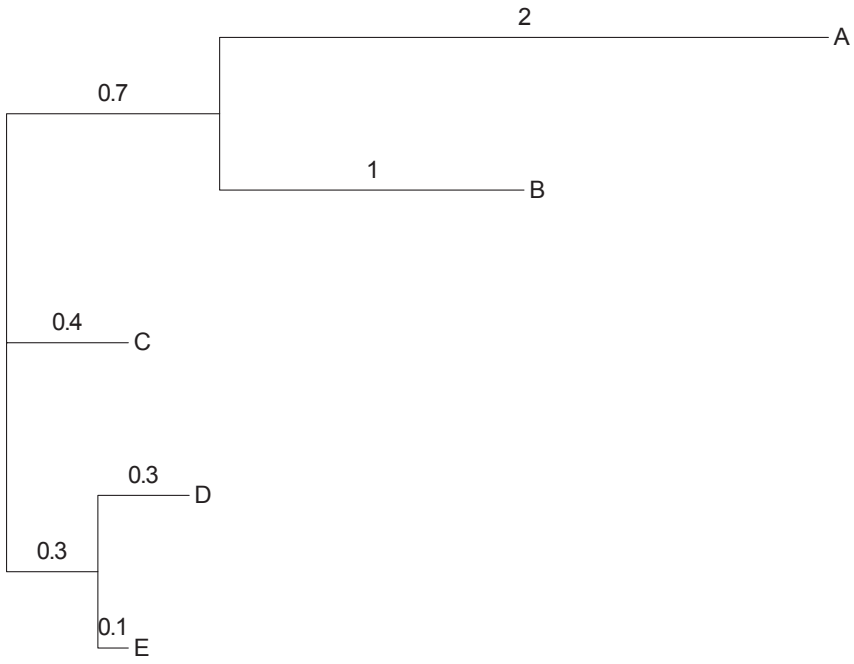
2.2.1 Normality of the pairwise distance estimates

A standard result is that the estimated pairwise distances have large-sample normal distributions (Susko 2003). In this section, we examine the distribution of the estimated pairwise distances to study how reasonable this approximation is. There are 40 different distances under each of the six "simulation-analysis" model pairs and Q-Q plots of these 40 distances all show similar patterns, so we only present Q-Q plots for six of them. The six distances are chosen to give a full range of true pairwise distances, and they are distance DE of tree 2 with length 0.4, CD of tree 1 with length 1.5, AB of tree 3 with length 2, BD of tree 4 with length 2, CE of tree 4 with length 2.2 and AD of tree 2 with length 3.3. The Q-Q plots are shown in the Figures 2.2-2.7.

From these Q-Q plots we find that JC69-JC69 model pairs exhibit the most serious skewness, This is because JC69 is too simple to incorporate all the variation in data. The distances in JC69 can be explicitly calculated as $-\frac{3}{4}\log(\frac{4}{3}(\hat{p}-\frac{1}{4}))$, where $\hat{p}$ is the proportion of sites with two nucleotides the same. Since $\hat{p}$ is approximately normally distributed, we expect the estimated pairwise distances to be normal only when $\hat{p}$ is much larger than $\frac{1}{4}$, where the logrithm function can be better approximated by a

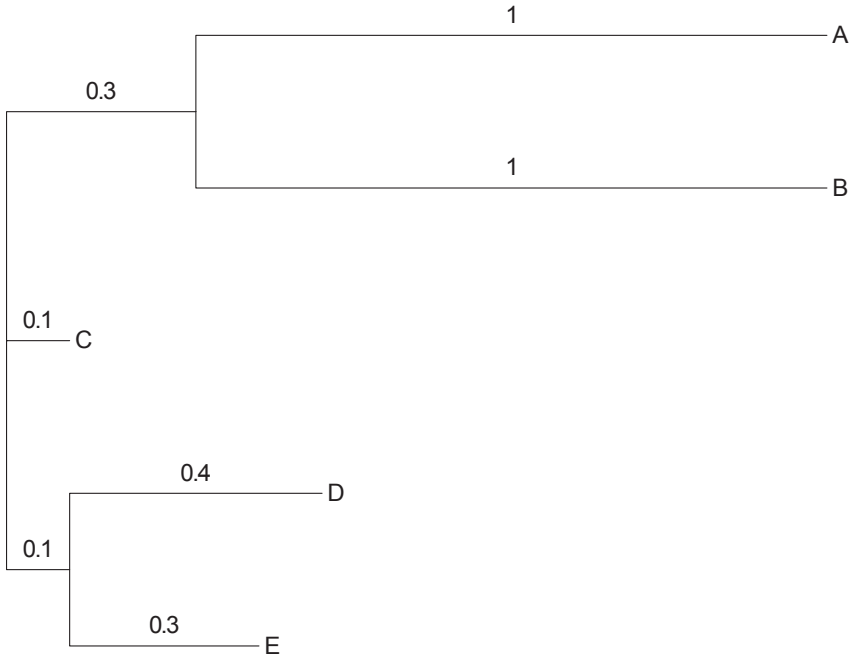


(a) Tree 1

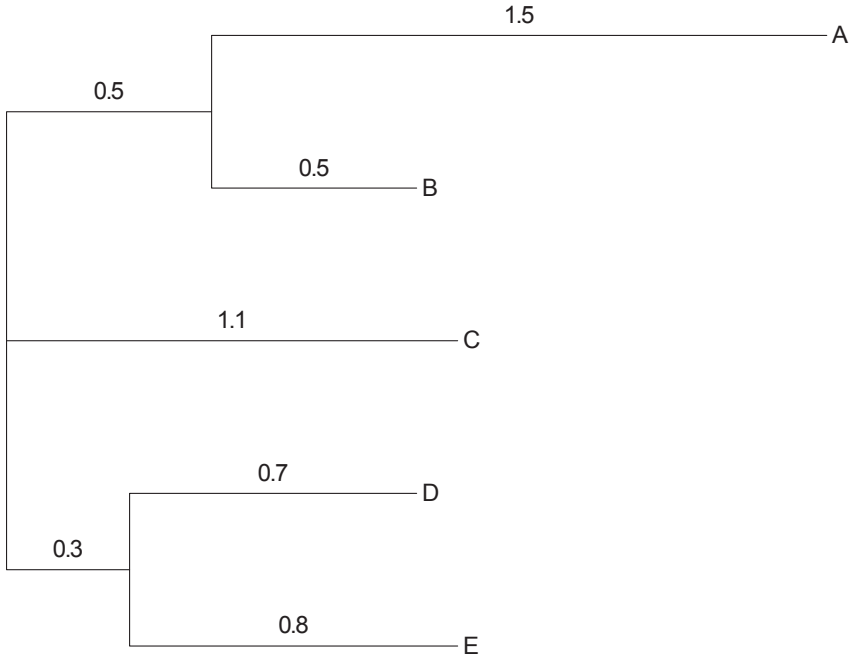


(b) Tree 2

Figure 2.1: Tree 1 and tree 2 used for simulation



(a) Tree 3



(b) Tree 4

Figure 2.2: Tree 3 and tree 4 used for simulation

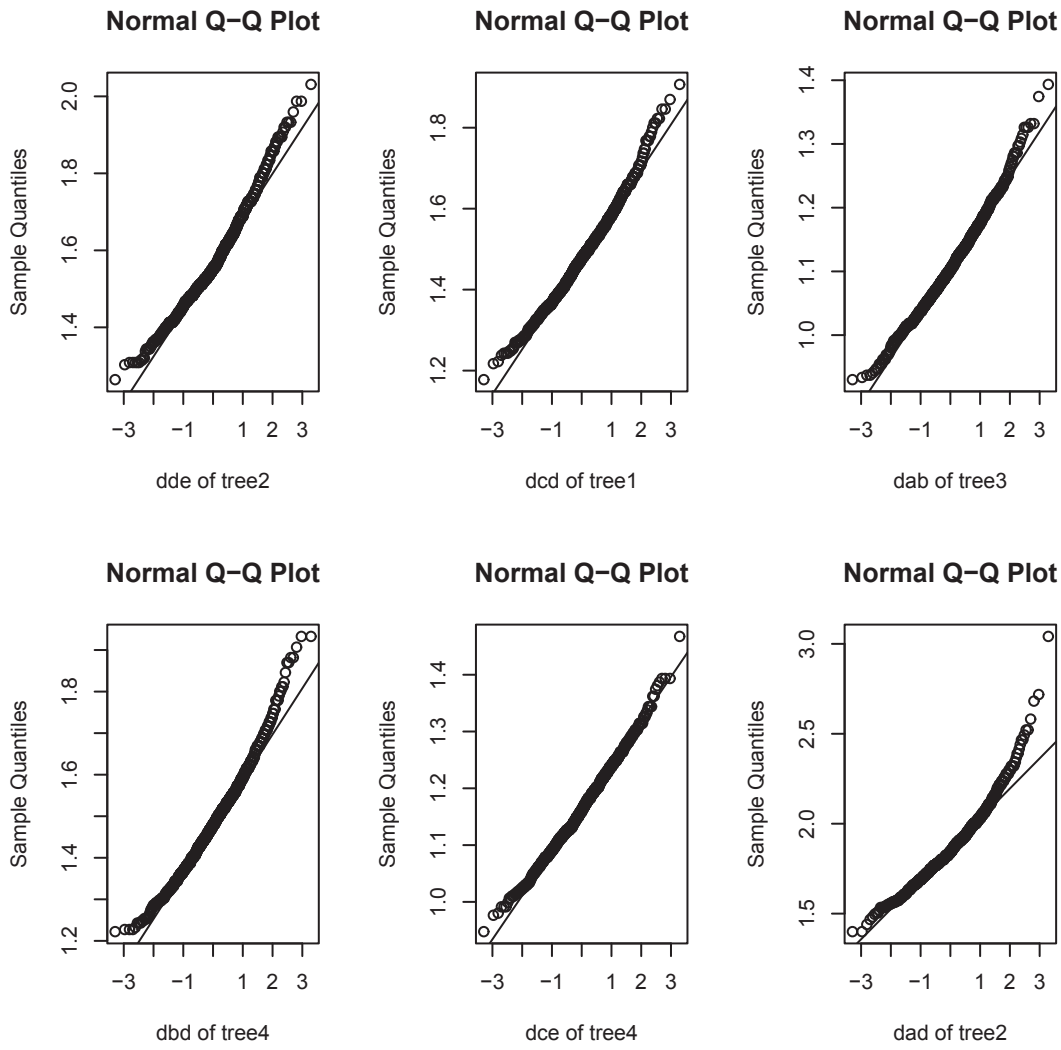


Figure 2.3: GTR-JC69 Q-Q plot

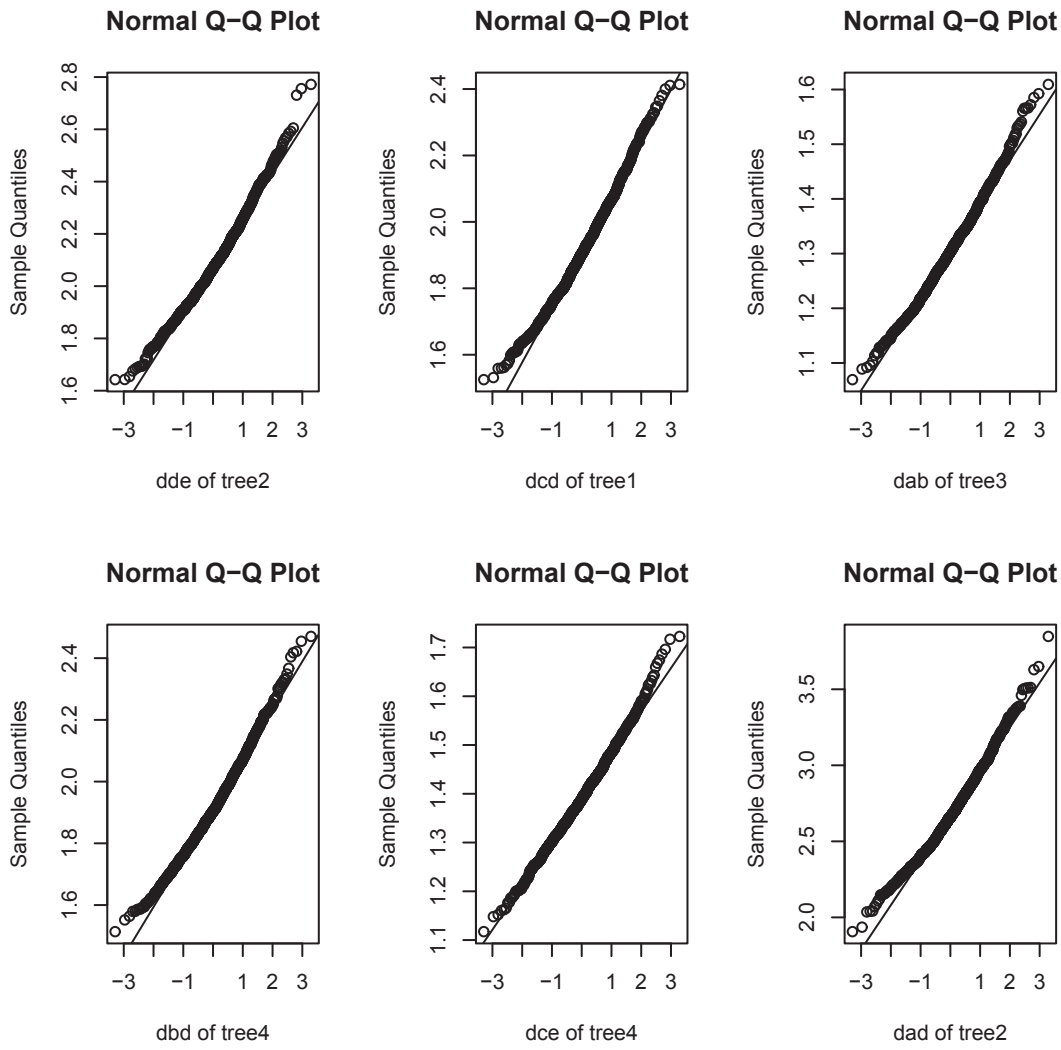


Figure 2.4: GTR-F84 Q-Q plot

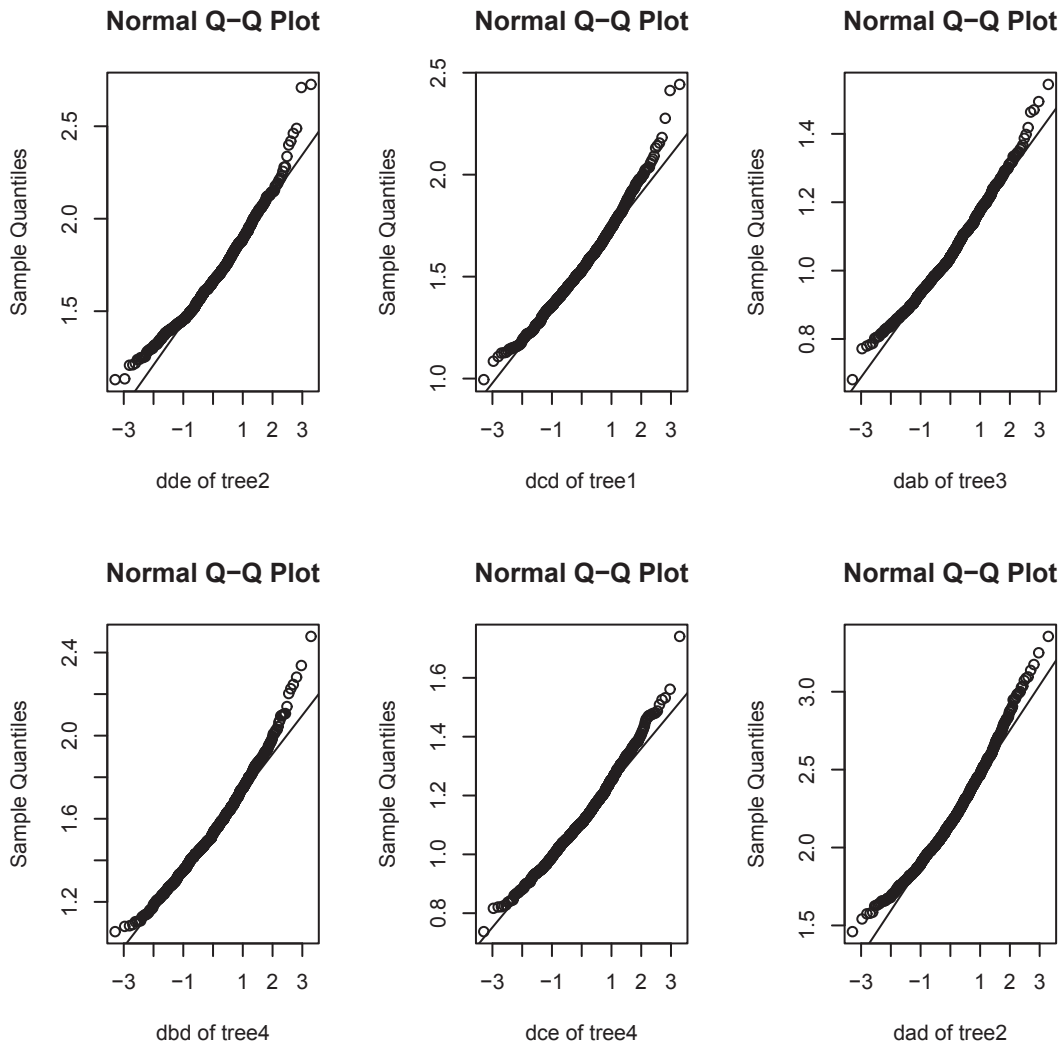


Figure 2.5: GTR+Gamma-F84 Q-Q plot

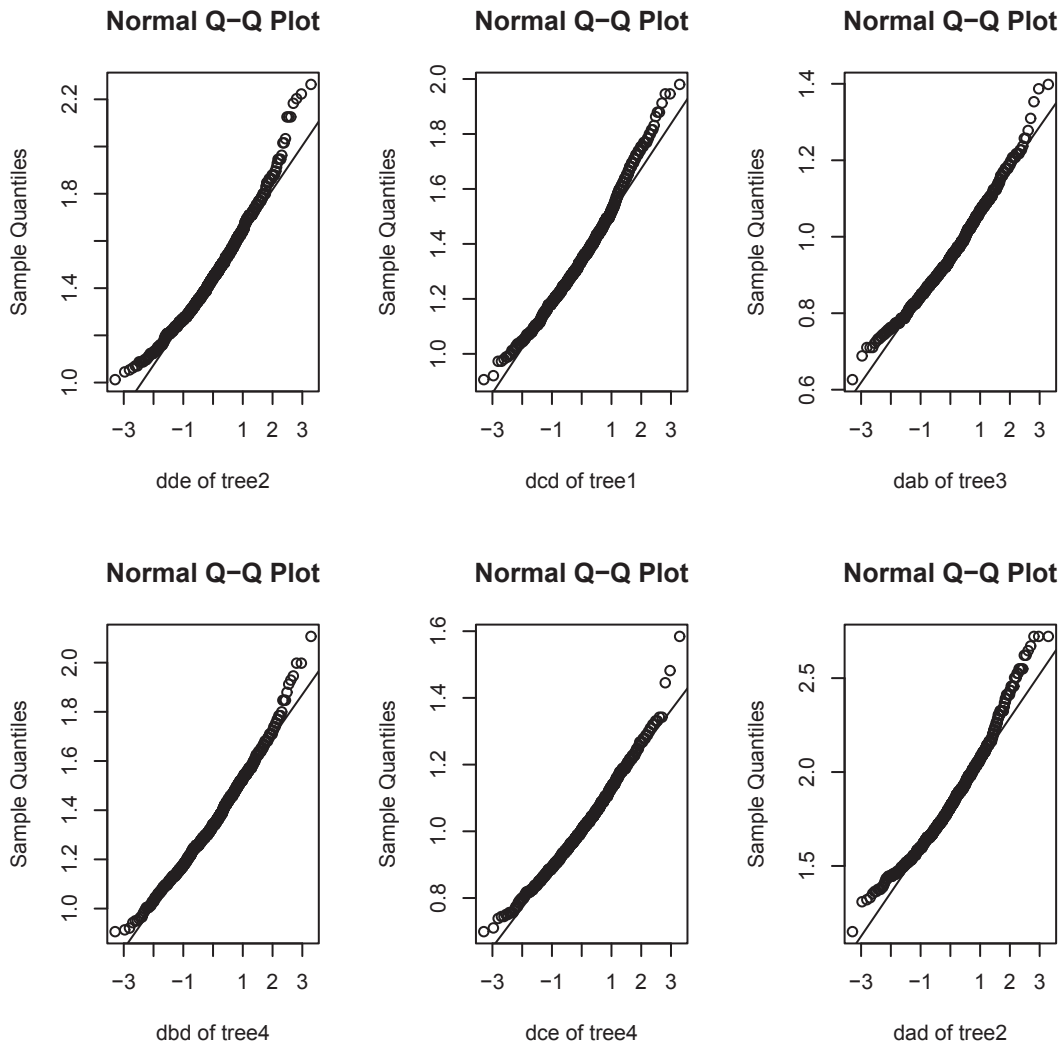


Figure 2.6: GTR+Gamma-JC69 Q-Q plot

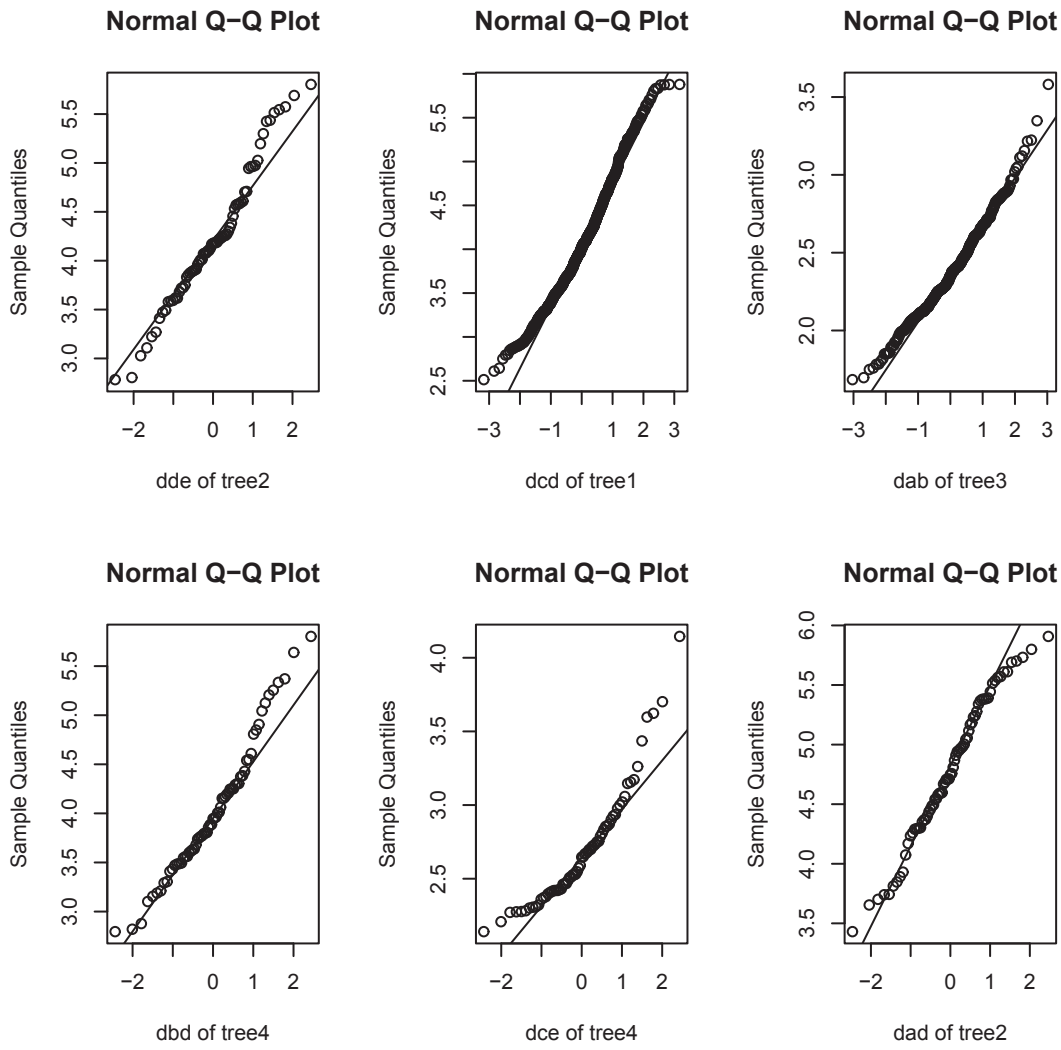


Figure 2.7: JC69-F84 Q-Q plot

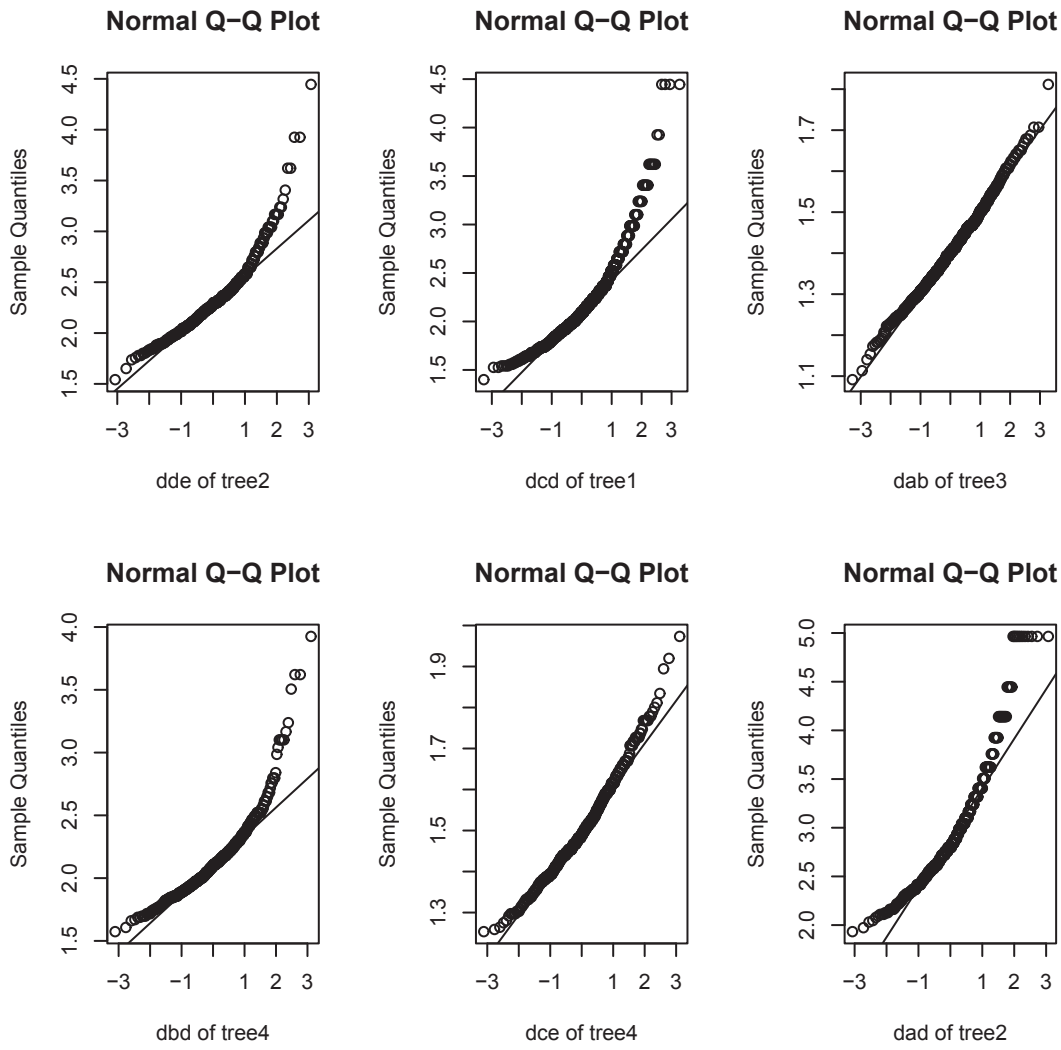


Figure 2.8: JC69-JC69 Q-Q plot

linear function of $\hat{p}$. This is possibly the case for the pairwise distance estimate AB of tree 3 under JC69-JC69 in Figure 2.7.

JC69-F84 in Figure 2.6 also shows some skewness, but note that JC69 is a model nested within F84, the skewness is less severe in JC69-F84 than those from JC69-JC69. For GTR-F84 in Figure 2.3, although there is model msspecification problem here, the pairwise distances follow normal curves very well. Some distances are not very normal for GTR-JC69 in Figure 2.2, such as distance AD of tree 2, again this is because JC69 is an oversimplified model. The Q-Q plots for GTR+Γ-JC69 and GTR+Γ-F84 show that the normality assumption works reasonably well for these model pairs.

It was found that under many model misspecification scenarios the estimated distances satisfy the normality property. But this is not guaranteed to be true, especially when the true pairwise distance is long. In this case the estimated pairwise distances can sometimes diverge, and the normal approximation fails badly. The Q-Q plot in Figure 2.8 is an example of this. It is for the model pair GTR-HKY85, where the true pairwise distance is 1.86 but the pairwise distances are output as 9 in many cases. The reason for this is that in some cases the sequences are divergent, meaning the formula for the distance does not define a value. PAML indicates this situation by outputing 9 in this case.

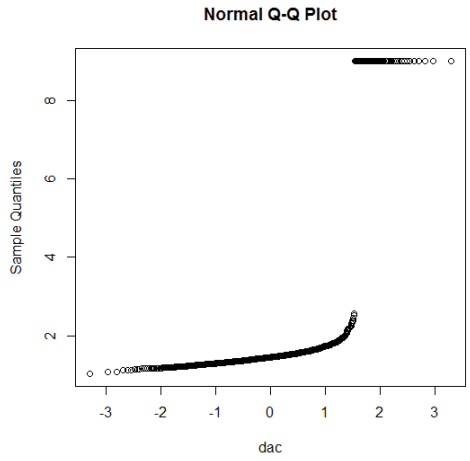


Figure 2.9: Q-Q norm plot of the longest pairwise distance in a hard tree

2.2.2 Infinite pairwise distance estimates

Undefined pairwise distances can influence statistical analyses. For example, statistics like variance, bias and mean squared errors of pairwise distance estimators are infinite, and distance-based methods for estimating trees do not have methods to deal with this. In practice, software packages use truncated values, and this can impair our ability to reconstruct the tree. We study how frequently this problem arises in typical situations. For a variety of model pairs and trees, we simulated 1000 replicates with sequence length 500, and counted how many times all the pairwise distance estimates were finite. The results are in Table 2.1.

	Model pairs					
Tree	GTR-F84	GTR-JC69	GTR+Γ-F84	GTR+Γ-JC69	JC69-F84	JC69-JC69
Tree 1	998	997	993	990	645	597
Tree 2	991	989	972	964	570	325
Tree 3	1000	1000	980	989	737	694
Tree 4	964	987	942	962	342	396

Table 2.1: Number of times (out of 1000) all pairwise distance estimates are finite (sequence length 500)

2.2.3 Bias, Variance and MSE of the pairwise distances

MSE is a good statistic to measure how much an estimate deviates from the true value. Recall that MSE = bias^2 + Variance. The previous section demonstrated that the normal approximation of the pairwise distance estimates is valid for many model misspecifications. In this section, we will find out how their bias, variance and MSE change under various model misspecifications.

Figures 2.9-2.14 show the MSEs, Vars and squared biases of the estimated pairwise distances versus the true pairwise distances for each tree under each model pair. For all trees, the MSEs of pairwise distance estimates always increase as the true pairwise distances increase. Under the model misspecification when the simulation models are GTR and GTR+Γ, all of the plots show that the squared biases are larger than the variances of the distance estimates, much larger in many cases. In some cases the variances are too small to be displayed in the plots. Comparing GTR-F84 and GTR-JC69 plots, we find that more serious model misspecification can lead to larger

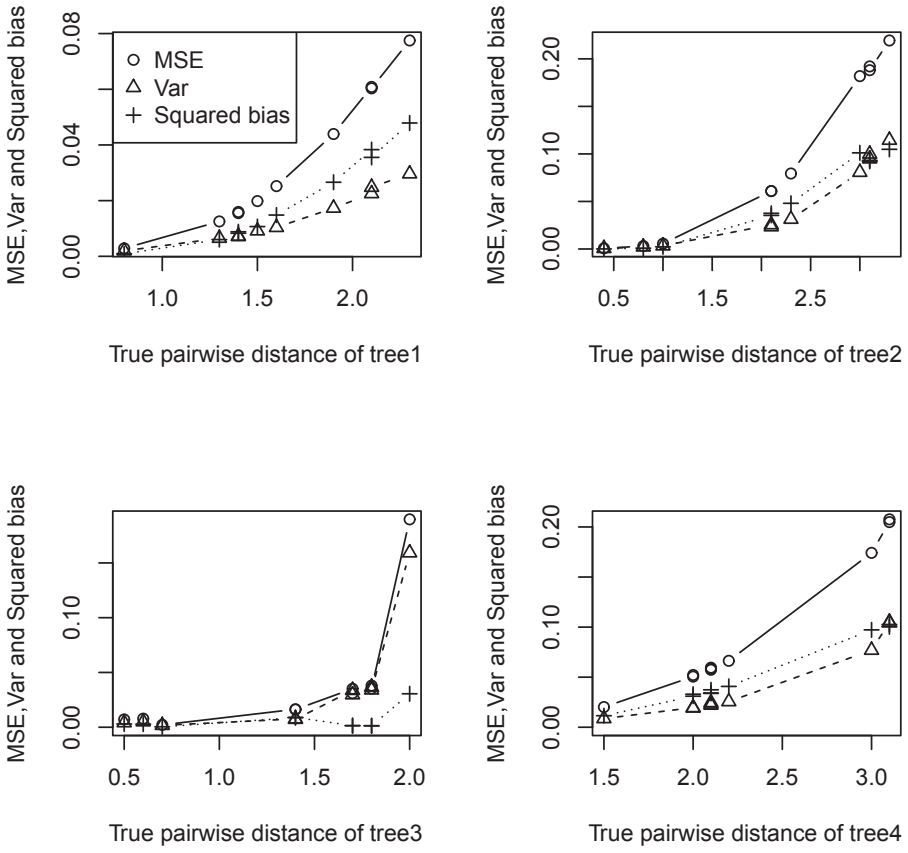


Figure 2.10: MSE,Var and Squared bias of GTR-F84

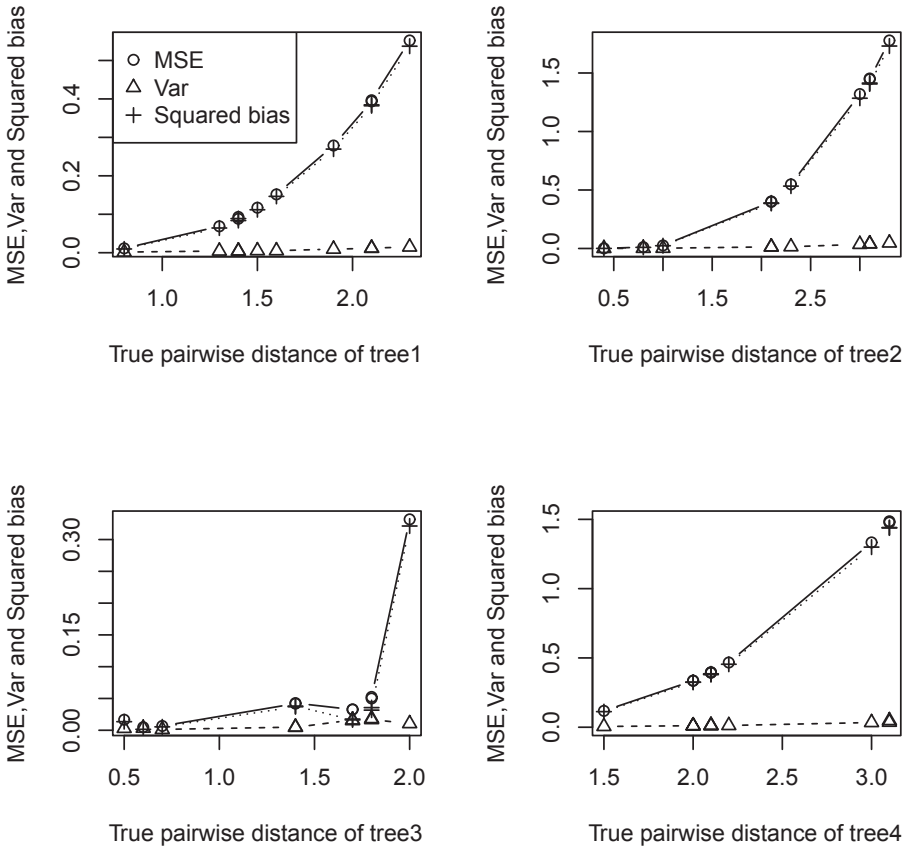


Figure 2.11: MSE,Var and Squared bias of GTR-JC69

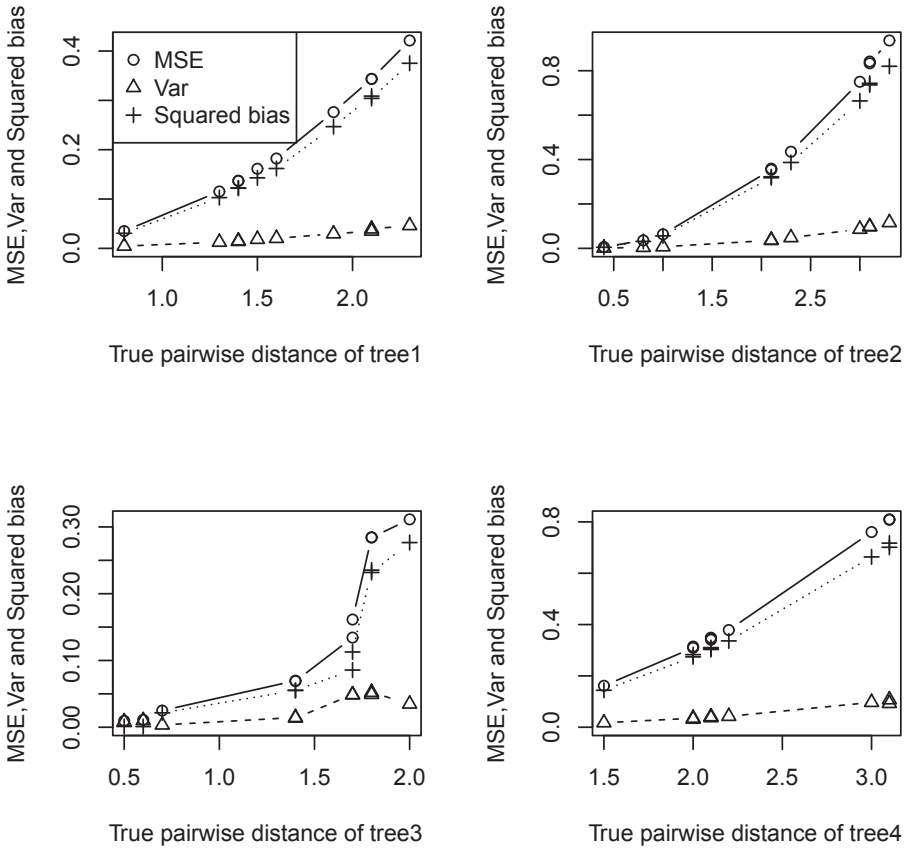


Figure 2.12: MSE,Var and Squared bias of GTR+Γ-F84

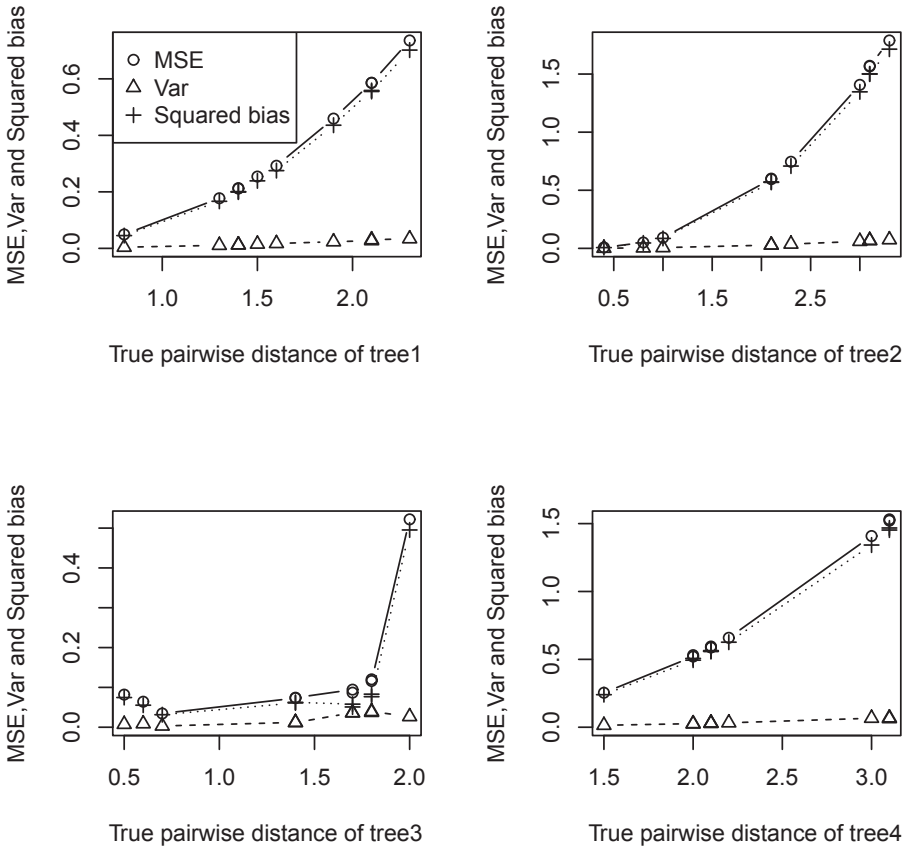


Figure 2.13: MSE,Var and Squared bias of GTR+Γ-JC69

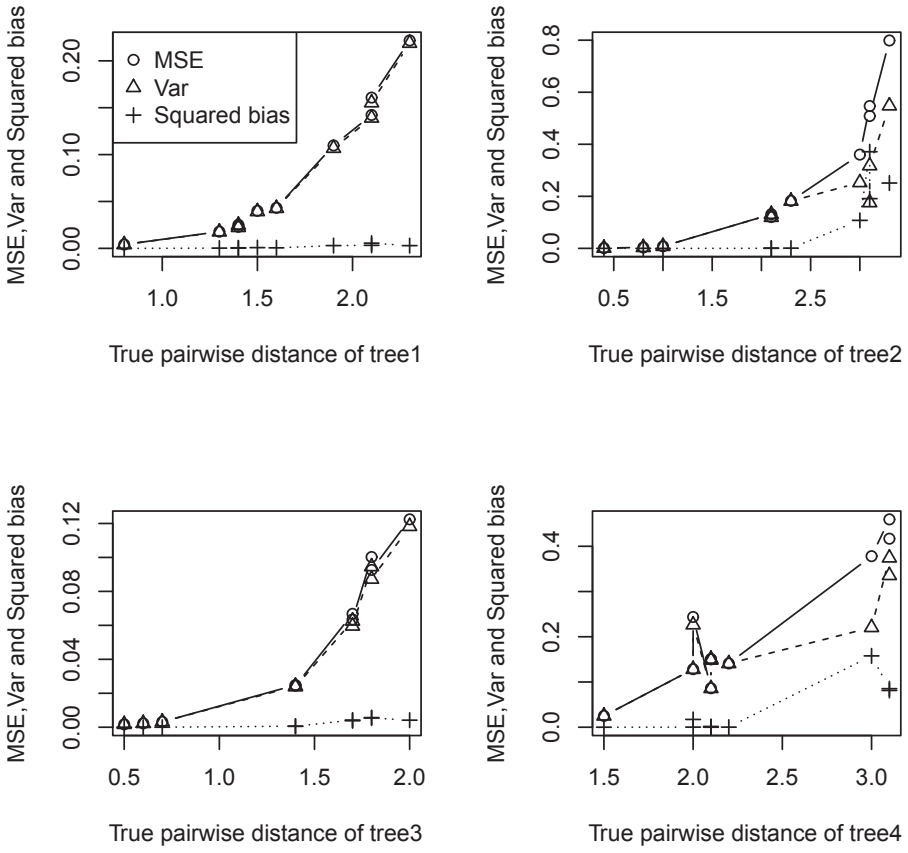


Figure 2.14: MSE,Var and Squared bias of JC69-F84

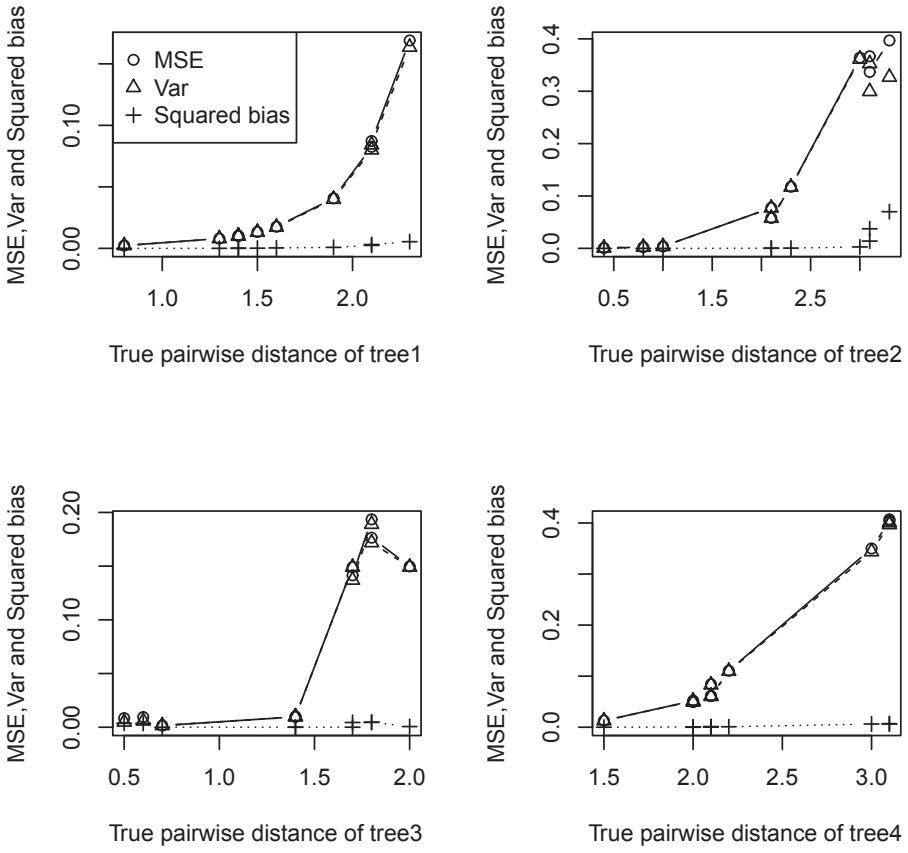


Figure 2.15: MSE,Var and Squared bias of JC69-JC69

bias. For example, in tree 2 the largest MSE and squared bias are approximately 0.2 in the GTR-F84 scenario, they are around 1.5 in the GTR-JC69 scenario. On the other hand, both JC69-JC69 and JC69-F84 plots (Figure 2.13 and 2.14) show that the biases are smaller than the variances. JC69-JC69 is without model misspecification and JC69-F84 is using an over-adequate model, thus the biases are smaller, but variances are relatively much larger.

2.2.4 Relationship between standard deviations (SD) and means of pairwise distance estimates

Some methods, such as weighted least squares (see Chapter 3), depend on variance assumptions for the pairwise distance estimates. In this section, we study how standard deviations are related to the mean for pairwise distance estimates. The plots are very similar under various model misspecifications, thus we only show the GTR-F84 case.

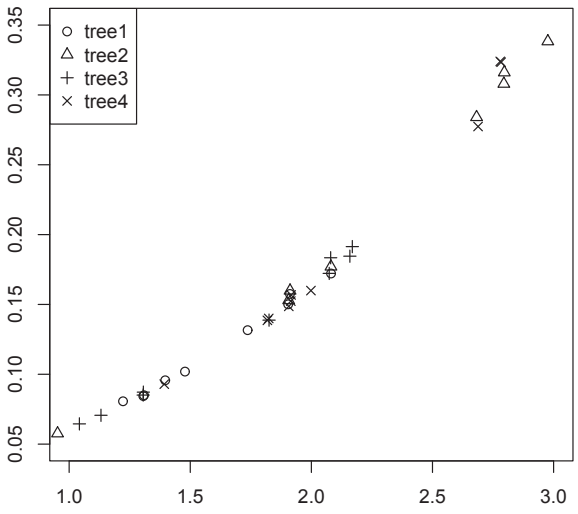


Figure 2.16: SDs against means of pairwise distance estimates, under GTR-F84

From Figure 2.15, we can see that the points from four different trees all appear to lie on a power curve, possibly quadratic curve. This is consistent with the weights often used in practice for weighted least squares methods.

2.2.5 Conclusion

When data are simulated under GTR and GTR+Γ, and the analysis model is misspecified, the pairwise distance estimates follow the normal distribution quite well. However, when the data are simulated under JC69, some of estimates do not follow the normal distribution. Model misspecifications lead to biased estimates for pairwise distances, and more severe misspecification, leads to larger bias. Without misspecification, the biases are very small, as expected. The SDs of pairwise distance estimates appear to have a power-law relationship with the means of pairwise distances estimates. Further study may be needed to confirm this relationship.

Chapter 3

Comparison of Balanced Minimum Evolution and Weighted Least Squares methods

Reconstructing the tree topology by distance-based methods requires the pairwise distance estimates first. The pairwise distance estimates may not be very accurate due to the model misspecification problems. Thus it is necessary to compare the effectiveness of some distance-based methods in tree topology estimates under model misspecifications. There are many distance-based methods developed so far, each with its own merits and drawbacks. For example, Neighbor-Joining (NJ) (Saitou and Nei 1987) method is fast at reconstructing the tree, while Weighted Least Squares (WLS) method (Fitch and Margoliash 1967) has the advantage in accuracy. The balanced minimum evolution (BME) method (Desper and Gascuel, 2002) is both fast and accurate. In this chapter, we first review the WLS and BME methods, then compare these two methods according to their tree topology estimate accuracies based on simulations.

3.1 A review of WLS method

We start by describing the ordinary least square (OLS) method, which is the foundation of the WLS method. Denote the tree being studied by T, and we have a distance estimate matrix $\Delta = (\delta_{ij})$, where Δ_{ij} is the pairwise distance estimate between taxa i and j, we will call Δ the *observed distances*. Let the $D = (d_{ij})$ be the matrix consisting of predicted distances d_{ij}'s, where the d_{ij} equals to the sum of all branch lengths between taxa i and j under T. Let $L = (l_k)$ be a vector representing the branch lengths of T, where l_k is the length of branch k. We use matrix $X = (x_{ij,k})$ to describe the topology of T: $x_{ij,k} = 1$ if branch k lies on the path between taxa i and j, and $x_{ij,k} = 0$ otherwise. The OLS minimizes sum of the squared differences

between observed and predicted distances.

$$\sum_{i=1}^{n}\sum_{j=1}^{n}(\delta_{ij} - d_{ij})^2$$

where n is the number of taxa. In matrix notation, the above criterion can be written as:

$$(XL - \Delta)^T(XL - \Delta)$$

The OLS solution to the branch length estimates is

$$\hat{L} = (X^T X)^{-1} X^T \Delta$$

However, the OLS method implicitly assumes that δ_{ij}'s are independent and have the same variance, neither of which is true. The WLS method minimizes the following:

$$(XL - \Delta)^T W^{-1}(XL - \Delta),$$

where W is a diagonal matrix, with its element w_{ij} proportional to the variance δ_{ij}. In practice, it is often assumed to be δ_{ij}^2 (Fitch and Margoliash 1967, Felsenstein 1997). The OLS method chooses the tree with the smallest sum of squared residuals while the WLS method prefers the tree with the smallest weighted sum of squared residuals (WSSR). The WLS estimates for L is:

$$\hat{L} = (X^T W^{-1} X)^{-1} X^T W^{-1} \Delta,$$

This diagonal matrix W does not contain any covariances of δ_{ij}'s.

In the generalized least squares (GLS) approach (Bulmer 1991; Susko 2003), the full variance-covariance matrix estimate of δ_{ij}'s is used. Incorporating all the covariances in the matrix increases the model complexity dramatically, leading to over-parametrized model. Susko (2011) has shown that WLS usually outperforms GLS at tree topology estimation in a wide range of typical situations.

3.2 A review of BME method

Desper and Gascuel (2004) demonstrated that BME is essentially a weighted least squares version of the minimum evolution (ME) method (Kidd and Sgaramella-Zonta,

1971; Rzhetsky and Nei 1993). ME is one of the basic principles in phylogenetic inference. ME selects the tree topology with the shortest length. The tree length is the sum of all branch lengths, which can be calculated from the pairwise distances. Denote the tree length of T as $S(T)$, $S(T)$ can be estimated using WLS as:

$$\hat{S}(T) = \mathbf{1}^T (X^T W^{-1} X)^{-1} X^T W^{-1} \Delta$$

where $\mathbf{1}$ is a vector of 1's. However, computing the tree length estimate in the above matrix form is computationally expensive, even though Gascuel (1997) and Bryant and Waddell (1998) have improved the computational speed for OLS, WLS and GLS greatly.

BME (Desper and Gascuel 2002) adopts another approach to estimate the tree length. To illustrate the idea, consider the branch configurations in Figure 3.1.

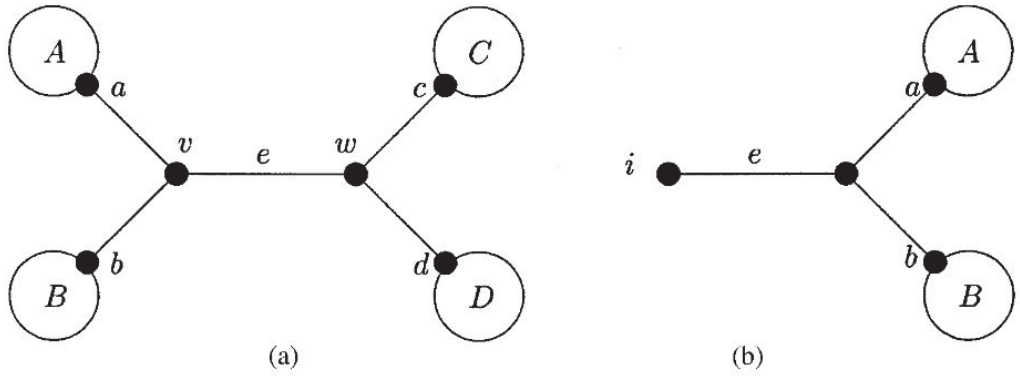


Figure 3.1: (a) for e an internal edge, (b) for e an external edge.

From Figure 3.1, when e is an internal branch, its estimate is

$$\hat{l}(e) = \frac{1}{4}(\delta^T_{AC} + \delta^T_{BD} + \delta^T_{AD} + \delta^T_{BC}) - \frac{1}{2}(\delta^T_{AB} + \delta^T_{CD})$$

and when e is an external branch,

$$\hat{l}(e) = \frac{1}{2}(\delta^T_{iA} + \delta^T_{iB} - \delta^T_{AB})$$

where δ^T_{AB} is the average distance between two subtrees A and B. If both A and B contain only one taxa a and b, then

$$\delta^T_{AB} = \delta_{ab}$$

if one of them, say B, consists of two subtrees, B_1 and B_2, then

$$\delta_{AB}^{T} = \frac{1}{2}(\delta_{AB_1}^{T} + \delta_{AB_2}^{T})$$

Pauplin (2000) showed that in this framework, the tree length estimate can be written as the following:

$$\hat{S}(T) = \sum_{i,j} 2^{1-B_{ij}}\delta_{ij}$$

where B_{ij} is the number of branches connecting taxa i and j.

In searching the tree topology, the principle of BME is coupled with the nearest neighbor interchange (NNI). NNI works by picking up an internal branch then exchanging the two subtrees connected to that internal branch. We use Figure 3.1 to demonstrate this. If applying the NNI to swap the subtrees B and C, we generate an alternative tree T^A, then the tree length difference between T and T^A:

$$\hat{S}(T) - \hat{S}(T^A) = \frac{1}{4}(\delta_{AB}^{T} + \delta_{CD}^{T} - (\delta_{AC}^{T} + \delta_{BD}^{T}))$$

Therefore BME circumvents the matrix manipulations involved in tree length estimation and improves the computational speed. Desper and Gascuel (2004) further demonstrated that BME has the following three properties:

1. Suppose the variance of δ_{ij} is proportional to $2^{B_{ij}}$ then the tree length estimate $\hat{S}(T)$ obtained from BME is identical to that from WLS. Moreover, it is the minimum variance tree length estimator.

2. BME is statistically consistent. Being consistent means when sequence length tends to infinity, with probability 1 the BME estimate of the tree is the true tree.

3. All the branch length estimates of all branches in tree T are non-negative, if T is a tree that is local minimum for a BNNI topology search.

The fast and accurate phylogeny reconstruction software (FastME) was developed based on the BME principle. The algorithm in FastME was refereed to as the BNNI algorithm. Vinh and Haeseler (2005) claimed that they found that BNNI boosts the topological accuracy of all distance-based methods.

3.3 Comparison of BME and WLS in tree topology estimates

Since both WLS and BME perform quite well in recovering the correct tree topology, we are curious about their performances in tree topology estimation under different situations. This section will compare BME and WLS based on simulations. Two kinds of simulations are designed. We firstly simulate the pairwise distances from normal distribution with various variance structures. This simulation design gives full advantages to the model assumptions of BME and WLS. We then simulate DNA sequence data based on some substitution models and observe the performances of these two methods under model misspecifcations.

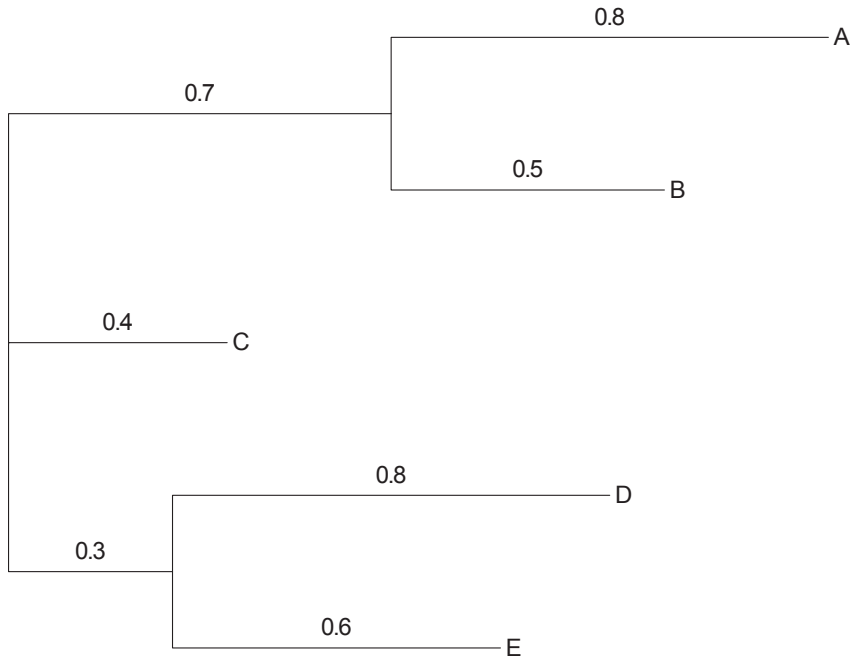
We use 5-taxa trees for simulation in this section. Five-taxa trees have a simple structure with only 15 different tree topologies, 10 pairwise distances and 7 branch lengths in total. BME selects the tree topology with the shortest tree length and WLS chooses the one with the smallest weighted sum of squared residuals (WSSR). It is not hard to find the best tree for either method for a 5-taxa tree, using an exhaustive search.

3.3.1 Simulation study based on pairwise distance simulations

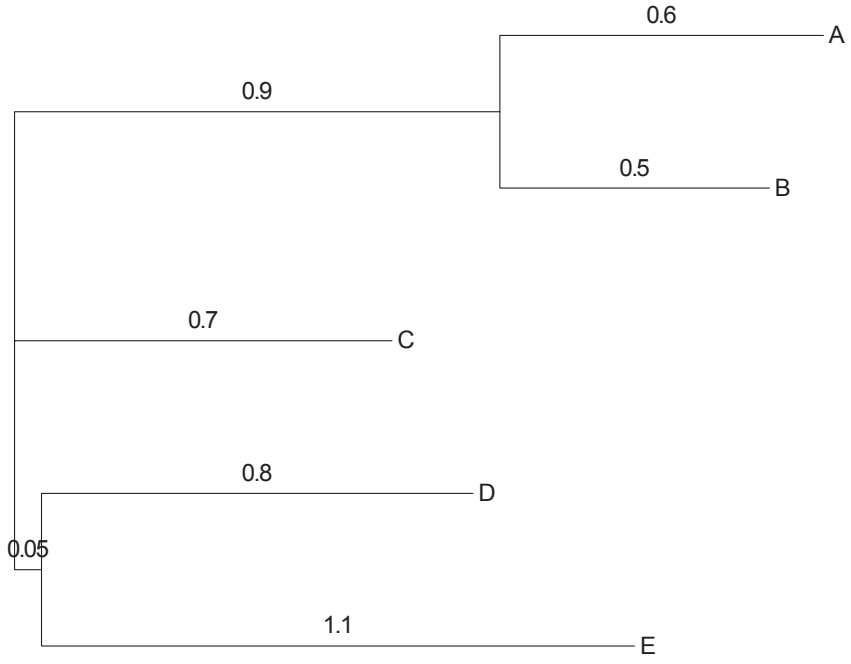
We consider two tree configurations as follows: the easy tree on the top panel of Figure 3.2, and the hard tree on the bottom panel of Figure 3.2. For the hard tree, the shortest internal branch is 0.05 and the longest external branch is 1.1, thus the longest branch is 22 times longer than the shortest branch. Each observed pairwise distance was independently simulated 100 times under normal distributions with mean equal to the true pairwise distance, and we apply three different variance structures in the simulations, namely:

1. $Var(\delta_{ij}) = k2^{B_{ij}}$.
2. $Var(\delta_{ij}) = kl_{ij}$, l_{ij} is the true pairwise distance between taxa i and j.
3. $Var(\delta_{ij}) = kl_{ij}^2$

The variance used by the WLS method to find the tree topology is $Var(\delta_{ij}) = kl_{ij}^2$, where k is a positive constant. We use R to calculate both the BME tree length estimates and WSSR of the WLS method, then count the frequencies that the BME



(a) Tree A



(b) Tree B

Figure 3.2: Tree A and tree B used for simulation

and the WLS choose the true tree. Note all true pairwise distances for both easy and hard trees happen to be greater than 1. We can sort the variance structure from large to small as: $k2^{B_{ij}} > kl_{ij}^2 > kl_{ij}$.

Simulation results and analysis on the easy tree

Table 3.1 lists the frequencies that different methods recover the true tree for different variance structures. The frequencies of both methods finding the true tree increase as the variance decreases. When the variance is very small, we find both methods choose the true tree 100% of the time. Moreover, under all these conditions, BME always performs slightly better than WLS with larger number of times selecting the correct tree.

		Variance structures used in simulation		
k values	Method	$k2^{B_{ij}}$	kl_{ij}^2	kl_{ij}
1/30	BME	51	84	92
	WLS	45	80	89
1/300	BME	99	100	100
	WLS	98	100	100
1/1000	BME	100	100	100
	WLS	100	100	100

Table 3.1: Frequencies of choosing the true tree for different k values and variance structures when the true tree is an easy tree. 100 replicates were simulated under each scenario

Simulation results and analysis on the hard tree

Table 3.2 lists the frequencies that different methods recover the true tree for different variance structures under another tree. Every frequency value in Table 3.2 is less than its corresponding value in Table 3.1 due to the fact that tree B is a hard tree. Neither method can 100% recover the true tree even when k is very small. When variance is large, both methods have difficulties in finding the true tree. Although tree B is a hard tree, we still find that BME performs better than WLS, which confirms the results from the simulations on the easy tree.

k values	Method	Variance structures used in simulation		
		$k2^{B_{ij}}$	kl_{ij}^2	kl_{ij}
1/30	BME	26	36	44
	WLS	25	35	42
1/300	BME	53	56	60
	WLS	52	56	59
1/1000	BME	61	71	90
	WLS	58	68	88

Table 3.2: Frequencies of choosing the true tree under different k values and variance structures when the true tree is a hard tree. 100 replicates were simulated under each scenario

Summary

For both the easy and the hard tree, under normal distribution simulation, the BME method is always slightly superior to the WLS method in true tree reconstruction. These results are obtained without applying substitution models to either simulations and analysis. In the next section, we will simulate data sets based on DNA models, analyze data with possible model misspecification problems, and examine the performances of both methods in finding the true tree.

3.3.2 Simulation study based on DNA sequence simulations

In this section, we simulate sequence data with GTR and GTR+Γ, and use F84 and JC69 models to analyze them. The trees we simulate are exactly the tree 2, tree 3 and tree 4 used in Chapter 2. The sequence lengths are 500 and 1000 respectively, and 1000 replicates are simulated under each scenario from Indelible1.03. Recall that the pairwise distance estimates follow a normal distribution very well in these cases. The frequencies of choosing true tree for each method under different scenarios are given in Tables 3.3 and 3.4.

Both methods increase their accuracy to find the true tree as the sequence length increases. This verifies that both methods are consistent. From Table 3.3, we can see that frequencies of finding the true tree by BME are usually larger than that by the WLS method under almost all different model pairs. From Table 3.4, BME's frequencies are sometimes lower than those from WLS, especially under tree 4, but their values are close. Thus these two methods are comparable in accuracy. Moreover,

		Model pairs			
Tree	Method	GTR-F84	GTR-JC69	GTR+Γ-F84	GTR+Γ-JC69
	BME	943	933	918	901
Tree2	WLS	809	705	919	900
	BME	900	672	755	637
Tree3	WLS	889	651	733	620
	BME	876	702	869	899
Tree4	WLS	865	690	857	907

Table 3.3: Frequencies of choosing the true tree with model misspecifications, sequence length is 500

		Model pairs			
Tree	Method	GTR-F84	GTR-JC69	GTR+Γ-F84	GTR+Γ-JC69
	BME	998	997	993	990
Tree2	WLS	999	997	993	989
	BME	999	986	977	956
Tree3	WLS	996	986	971	944
	BME	964	899	936	910
Tree4	WLS	967	907	946	901

Table 3.4: Frequencies of choosing the true tree with model misspecifications, sequence length is 1000

when sequence length is 1000, both methods have over 90% probability of finding the true tree, so there is little difference between these two methods in true tree topology reconstruction in this case. We can conclude that with model misspecifications, BME performs as well as or better than WLS in tree topology reconstruction.

3.4 Conclusion

BME is a powerful method in finding the true tree topology, it is fast and accurate. We find that under many scenarios, BME performs slightly better than WLS. The WLS method and some other least squares methods have been widely applied in phylogenetic inference, for example for building confidence regions. In the next chapter, we use BME to construct a confidence region.

Chapter 4

Constructing a confidence region using BME

We demonstrated in Chapter 3 that BME is a powerful method once reliable pairwise distance estimates are available, but sometimes it is difficult to determine the true tree. That motivates us, instead of recovering a single tree estimate, to try to determine a confidence region. Historically, construction of confidence regions is based on tests. Many test procedures have been proposed. For ML methods; such as the Shimodaria-Hasegawa (SH) test (Shimodaira and Hasegawa 1999), the SOWH test (Swofford et al. 1996, Goldman Anderson and Rodrigo 2000), the approximately unbiased (AU) test (Shimodaira 2002), the likelihood weight (LW) test (Strimmer and Rambaut 2001) and the single distribution nonparametric bootstrap (SDNB) test (Shi et al. 2005) etc. For distance-based methods; there are methods based on the WLS and GLS tests (Susko 2003). In this chapter we extend the idea of the SDNB test to the BME method to construct confidence regions. We then examine its performance on simulated data.

4.1 A review of the SDNB test and WLS and GLS tests

Let the tree topology being tested be denoted τ_0. The null hypothesis is

$$H_0 : \tau_0 \text{ is the true tree topology}$$

The branch lengths are not specified here but estimated during the test. The corresponding P-value is the probability that the test statistic is no less than the value observed. So given a significance level α, we can construct a $(1-\alpha)\times 100\%$ confidence region by testing. This confidence region is a set of topologies that we expect the true tree topology is in with probability at least $1-\alpha$. It includes τ_0 if H_0 is not rejected with significance level α. The coverage here is defined as probability that the confidence region contains the true tree topology. The size of the confidence region is the number of tree topologies it contains. Both the size and coverage are

evaluated by simulations and these two are important criteria to judge the test. In order to generate the distribution of the test statistic, nonparametric bootstrap was performed, since there is no closed-form for the distribution. Specifically, for DNA sequence data, we use a nonparametric bootstrap on sites to generate replicates, and the combinations of nucleotide types in sites are regarded as bootstrap replicates.

Most existing ML based tests apply the ML method to estimate the tree lengths. Suppose the log likelihoods of tree τ_i and the ML tree are denoted as l_i and l_{ML} respectively, then we have test statistics

$$\delta_1 = l_{ML} - l_1, \cdots, \delta_m = l_{ML} - l_m$$

SDNB test approximates the null distributions for all m tests by one distribution, generated by nonparametric bootstrap. Since the ML tree of the original data set is considered to represent the most probable evolutionary process, in the bootstrap, the ML tree of the original data set is used to replace the tested tree to calculate bootstrap statistics according to the bootstrap theory. The SDNB method follows the following procedures :

1. Generate nonparametric bootstrap replicate data sets.

2. Obtain the maximum log likelihood $l^*_{ML^*}$ and l^*_{ML} by reestimation. l^* denotes the log likelihood for the bootstrap data set. The subscript ML and ML^* are used to represent the ML tree τ_{ML} of the original data set and ML tree τ_{ML^*} of the replicate respectively.

3. Construct the distribution of $\delta_i = l_{ML} - l_i$ by calculating the difference between the ML tree of the replicate and the ML tree of the original data set, $\delta^* = l^*_{ML^*} - l^*_{ML}$. Compare the δ^* from the bootstrap to the test statistics from the original data sets.

4. For significance level α, if δ_i is less than the $100(1-\alpha)\%$th percentile of δ^*, then tree τ_i is included in the $100(1-\alpha)\%$ confidence region.

In this algorithm, only one bootstrap distribution is required, to test all potential tree topologies, so this is computationally easier than calculating one distribution for each test.

Distance-based methods are also employed in some tests. The GLS method (Susko 2003) used the following test statistic, which is exactly the same one as the WLS statistic described in Chapter 3.

$$(XL - \Delta)^T W^{-1} (XL - \Delta)$$

where $X = (x_{ij,k})$ is the tree topology matrix, L is the vector representing the branch lengths of the tree, and Δ is the matrix consisting of estimated pairwise distances. Matrix W is a full variance-covariance matrix in this case. Under the null hypothesis, the test statistic above approximately follows the chi-squared distribution with degrees of freedom equal to $T(T-1)/2 - (2T-3)$, where T is the number of taxa.

The WLS method (Czarna et al. 2006, Susko 2011) provides an alternative means of constructing the confidence region of tree topologies. The test statistic is very similar to that of GLS except that matrix W is not the full variance-covariance matrix, but only a diagonal matrix with variances of pairwise distance estimates, i.e WLS method ignores the covariances. The WLS method is superior to the GLS for stability of the test statistic. It is possible that the variance-covariance matrix in the GLS test statistic is singular, but the variances of pairwise distances are not close to 0, therefore the test statistic of the WLS method is better behaved. The real problem for the WLS method is that under the null hypothesis, the distribution of its test statistic is not available. Czarna et al. (2006) suggest assuming the independence of the pairwise distances. If this is true, then the WLS's test statistic approximately follows the same distribution as the GLS test. However, simulations (Susko 2011) indicate this usually is not true.

4.2 A new test for confidence region construction

The above tests have some shortcomings: it is computationally expensive to apply ML for the large taxa trees; GLS is also inconvenient because the covariances of pairwise distance estimates are poorly known; WLS has the disadvantage that the distribution of the test statistic is unknown. Susko (2011) gives a fast parametric bootstrap approach for approximating this distribution, but we do not pursue this further here.

In Chapter 3, we demonstrated that BME is powerful not only for its fast computation speed to calculate the tree length but also for its accuracy to find the true tree. We now extend the idea of SDNB test to BME for constructing confidence region.

Suppose the tree length of tree τ_i and the best tree found under the BME method are s_i and s_{BME} respectively, their difference, denoted as $d_i = s_i - s_{BME}$, is the test statistic for testing the null hypothesis H_0: τ_i is the true tree. The key steps of this algorithm are:

1. Generate nonparametric bootstrap replicate data sets.

2. Reestimate the tree length $s^*_{BME^*}$ and s^*_{BME} of each replicate bootstrap data set. s^* denotes the tree length estimate obtained by BME method for the bootstrap data set. BME and BME^* denote the best trees found by BME method from the original data set and the replicate data set, respectively.

3. Create a distribution of test statistic by calculating the differences $d^* = s^*_{BME} - s^*_{BME^*}$, which is the differences of tree lengths between the best tree of the original data set and the best tree of the replicate data sets. Then for significance level α, obtain its $100(1-\alpha)\%$ percentile value, which is denoted as c_α here.

4. If d_i is less than c_α, then tree τ_i is included in the $100(1-\alpha)\%$ confidence region. Note that the BME tree of the original data set must be in this confidence region since $d_{BME} = s_{BME} - s_{BME} = 0$. Therefore the size of this confidence region is at least 1.

5. When the number of taxa is small (eg. 5), it is possible to search the whole tree space, and construct the confidence region by testing each tree. However, when the number of taxa is large, the tree space is too large for each tree to be tested. We will apply NNI to construct the confidence region. We create a list L of trees in the confidence region. As a first attempt, we start with a list consisting of just T_{BME}, and inductively create the full list as follows:

   (a) For each tree in this list, find all trees not already considered that can be reached from this tree by performing a single NNI operation on one (internal) branch. Calculate the lengths of each of these trees, and determine

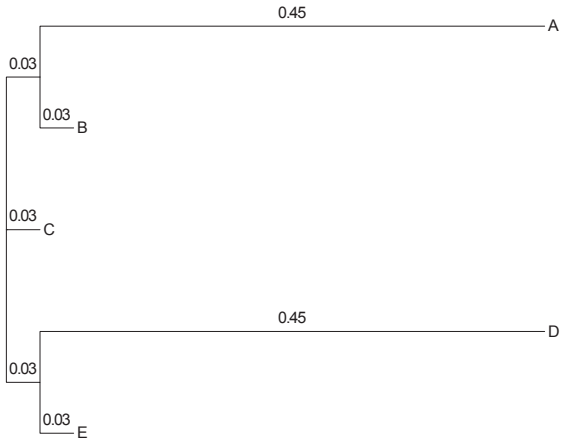
which should be added to L by testing whether the tree length is less than $S_{BME} + c_\alpha$

(b) Continue until step (a) does not produce new trees in the list L .

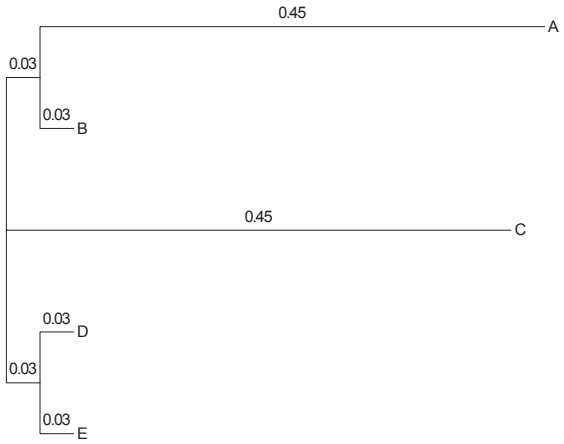
The number of trees in the list L is the size of the confidence region. The coverage is evaluated by whether the true tree is in the confidence region. The trouble with this approach is that if the trees near to T_{BME} are not in the confidence region, then the NNI search will not search many trees, and as a result we may miss the true tree. To correct for this, we can modify the algorithm to a two-pass algorithm, where in the first pass, we perform a broad search using the same algorithm but with a different cut-off c^* instead of c_α. We then test each tree in this large list L compared to the test statistic c_α.

4.3 Simulation design and analysis

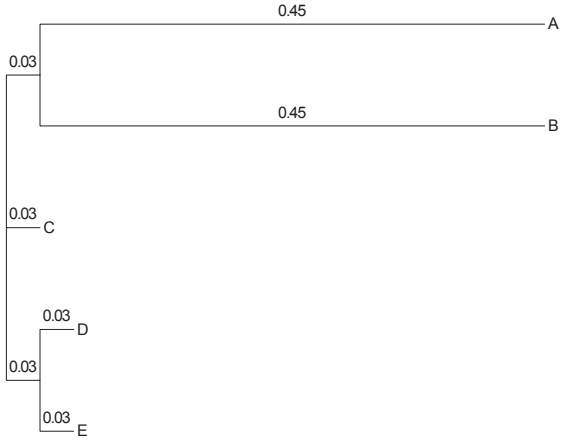
We consider trees with both small and large number of taxa. All the trees were simulated and analyzed under HKY, so there is no model misspecification involved. For the small number of taxa, the parameters of the HKY model were set as $\pi_a = 0.18, \pi_c = 0.33, \pi_g = 0.26, \pi_t = 0.23$ with ratio of transition/transversion set as 2.93 as in (Shi et al. 2005). For the large number of taxa, the parameters were set as $\pi_a = 0.37, \pi_c = 0.24, \pi_g = 0.12, \pi_t = 0.27$ as in (Zwickl and Hillis, 2002) with the same transition/transversion ratio used as for small number of taxa. Each data set was simulated 100 times using Indelible1.03 and we used seqBoot from the PHYLIP package to generate 1000 nonparametric bootstrap replicates of each original data set. PAML (Yang 1994) was used to estimate the pairwise distances for trees of small numbers of taxa, and we applied BME to estimate their tree lengths. For large numbers of taxa, we directly used the FastME program to obtain their tree length estimates. For both small and large number of taxa, we set $\alpha = 0.05$, i.e we only examine the 95% confidence region for all simulation studies. The simulated data sets were all of length 1000 nucleotides.



(a) Tree 1



(b) Tree 2



(c) Tree 3

Figure 4.1: Tree 1, tree 2 and tree 3 used for simulation

4.3.1 Small number of taxa

We tested three different trees of 5-taxa in Figure 4.1 with their branch lengths shown below, these are from Shi et al. (2005) with all the branch lengths reduced to half the original values. The reason for reducing the branch lengths is that estimating large pairwise distances is not possible for reasonable sequence lengths. We used the average size to measure the size of confidence region. The results of our simulations are shown below

Tree	Diameter	Coverage	Average size	Standard deviation
1	0.96	100	1.2	0.4
2	0.93	100	2.8	0.79
3	0.9	98	3.4	1.62

Table 4.1: Coverage and average size of confidence region for trees with 5-taxa

In Shi et al (2005), they introduced percentage of time that the ML and the true trees (PMLT) are the same to measure how difficult the phylogenetic analysis is. The smaller PMLT, the harder the tree analysis. For tree 1 and tree 2, their PMLT for trees of the original size are 100 and 97 respectively; for tree 3 it is 68, so tree 3 is a relatively harder tree. Our goal is to construct a 95% confidence region of small size with coverage is also at least 95 %. Our results are shown in Table 4.1. The average size of confidence regions is around 3 in Shi et al (2005) which was calculated over all 12 model trees. Our results look comparable to theirs.

4.3.2 Large number of taxa

When the number of the taxa becomes large, we depend on the tree-searching method to choose the candidate trees, NNI was applied both to search for the optimal tree, and to search for candidate trees. We examine the performance of this test based on a 15-taxa tree, also from Shi et al (2005), with its topology shown below. There are over 7.9×10^{12} different tree topologies for 15-taxa tree. For this analysis, we used $c^* = 1.5c_{0.05}$. It is important to note that our results are not comparable to (Shi et al. 2005) because our trees are shorter and they did not perform a search of tree space, but instead tested only a set of 100 trees.

Results are: average size is 392.09 with the standard deviation 3.17 and the coverage is 100%, there are 545 trees were included in the first cut on average.

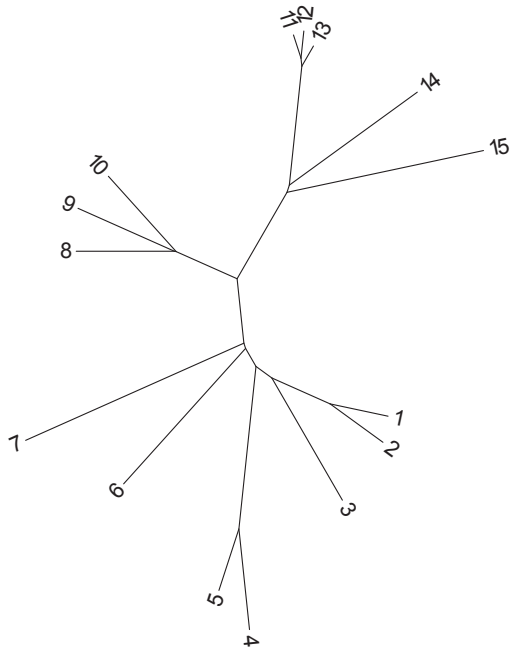


Figure 4.2: The structure of 15 taxa tree for simulation

4.4 Conclusion

We proposed a new test combining with the BME method and the SDNB test. Unlike implementations of other tests, our test searches the tree space to construct the confidence region. For large taxa trees, since BME is very fast for estimating the tree our method has potential to provide a fast method for constructing a good confidence interval. The current method for computing the distribution of the test statistic is computationally expensive, so further work is needed to find this distribution more efficiently. This is a very feasible problem, and we are optimistic that this test can lead to a very fast and effective method for constructing a confidence region. Further work is also needed to find the best value for the cut-off c^*.

Chapter 5

Conclusion

Reconstructing the tree topology based on distance method needs to estimate their pairwise distances first. Usually, without model misspecification, the pairwise distances estimates follow the normal curve well, but under the model misspecified cases, the normality assumption can not be preserved. Also if the data is simulated under a simple model but analyzed under a more complicated one, such as JC69-GTR, the squared biases of pairwise distance estimates are expected to be smaller than the corresponding variances, and the results are opposite for "advanced-simple" model pairs. Under various variance structures, and under different model pairs, BME performs at least as well as WLS at choosing the tree topology. Moreover, BME computes much faster since it avoids the matrix manipulations needed in other methods. This motivates us to apply the BME method to construct a confidence region. We have given a method for constructing a confidence region based on BME. This method has shown good results in terms of both coverage and size of confidence region.

5.1 Future work

This new method for finding a confidence region is promising, but there is a lot of fine-tuning that can be done to improve the results. Probably the most important fine-tuning is to find a faster way to estimate the distribution of the test statistic, one of the big advantages of BME is its speed, and the bootstrap loses this advantage. Given the usual assumptions about the distribution of pairwise distances. it should be feasible to estimate the distribution of the test statistics in a more computationally efficient way.

Bibliography

[1] David Bryant, Peter Waddell. 1998. "Rapid Evaluation of Least-Squares and Minimum-Evolution Criteria on Phylogenetic Trees." *Mol. Biol. Evol.* 15(10): 1346-1359.

[2] M. Bulmer. 1991. "Use of the method of generalized least squares in reconstructing phylogenies from sequence data." *Mol. Biol. Evol.* 8:868-883.

[3] Richard Desper, Oliver Gascuel. 2002. "Fast and Accurate Phylogeny Reconstruction Algorithms Based on the Minimum-Evolution Principle." *Journal of Computational Biology* 9(5) 687-705.

[4] Richard Desper, Oliver Gascuel. 2004. "Theoretical Foundation of the Balanced Minimum Evolution Method of Phylogenetic Inference and Its Relationship to Weighted Least-Squares Tree Fitting." *Mol. Biol. Evol.* 21(3):587-598.

[5] 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).

[6] 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.

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

[8] J. Felsenstein. 1984. "Evolutionary trees from DNA sequences: a maximum likelihood approach". *Journal of Molecular Evolution* 17 (6): 368-376.

[9] J. Felsenstein. 1989. PHYLIP Phylogeny Inference Package (Version 3.2). *Cladistics* 5, 164-16

[10] J. Felsenstein. 1997. "An alternating least-squares approach to inferring phylogenies from pairwise distances." *Syst. Biol.* 46, 101-111.

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

[12] W.M. Fitch and E. Margoliash. 1967. "Construction of phylogenetic trees." *Science* 155, 279-284

[13] O. Gascuel. 1997. "BIONJ: An improved version of the NJ algorithm based on a simple model of sequence data." *Mol. Biol. Evol.* 14, 685-695.

[14] N. Goldman, J. P. Anderson and A. G. Rodrigo. 2000. "Likelihood-based tests of topologies in phylogenetics." *Systematic Biology.* 49:652-670.

[15] 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.

[16] 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.

[17] P.J.Lockhart et al. 1992. "Substitutional Bias Confounds Inference of Cyanelle Origins from Sequence Data." *J. Mol. Evol* 34: 153-162.

[18] P.J.Lockhart et al. 1994. "Recovering Evolutionary Trees under a More Realistic Model of Sequence Evolution." *Mol. Biol. Evol.* 11(4):605-612.

[19] Y.Pauplin. 2000. "Direct Calculation of a Tree Length Using a Distance Matrix." *J.Mol.Evol.* 51:41-47

[20] A. Rzhetsky and M. Nei. 1993. "Theoretical Foundation of the Minimum-Evolution Method of Phylogenetic Inference." *Mol. Biol. Evol.* 10(5): 1073-1095

[21] N. Saitou and M. Nei. 1987. "The neighbor-joining method: A new method for reconstructing phylogenetic trees." *Mol. Biol. Evol.* 4, 406-425.

[22] S. Tavare. (1986). "Some Probabilistic and Statistical Problems in the Analysis of DNA Sequences". *Lectures on Mathematics in the Life Sciences* (American Mathematical Society) 17: 57-86

[23] X. Shi., H.Gu., E. Susko. and C. Field. 2005. "The Comparison of the Confidence Regions in Phylogeny." *Mol. Biol. Evol.* 22(11): 2285-2296.

[24] H. Shimodaira and M. Hasegawa. 1999. "Multiple Comparisons of Log-Likelihoods with Applications to Phylogenetic Inference". *Mol. Biol. Evol.* 16(8):1114-1116.

[25] H. Shimodaira. 2002. "An Approximately Unbiased Test of Phylogenetic Tree Selection". *Syst. Biol.* 51(3):492-508.

[26] K. Strimmer and A. Rambaut. 2001. "Inferring conidence sets of possibly misspecified gene trees." *Proc. R. Soc.* 269: 137-142.

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

[28] E. Susko. 2003. "Confidence Regions and Hypothesis Tests for Topologies Using Generalized Least Squares." *Mol. Biol. Evol.* 20(6):862-868.

[29] E. Susko, Y. Inagaki, and A.J. Roger (2004). "On inconsistency of the neighbour joining method and least squares estimation when distances are incorrectly specified." *Molecular Biology and Evolution*, 29:1629–1642.

[30] E. Susko. 2011. "Improved Least Squares Topology Testing and Estimation." *Systematic Biology*. 60:668–675.

[31] D. L. Swofford , G. J. Olsen, P. J. Waddel, and D. M. Hillis. 1996. Phylogenetic inference. Pages 407-514 in *Molecular systematics*, 2nd. ed. (D. M. Hillis, B. K. Mable, and C. Moritz, eds.). Sinauer, Sunderland, Massachusetts.

[32] L.S. Vihn, von Haeseler. 2005. "Shortest triplet clustering: reconstructing large phylogenies using representative sets." *BMC Bioinformatics* 6:92

[33] Ronald A. Van Den Bussche, J. Baker. Robert, John. P. Huelsenbeck and David. M. Hillis,. 1998. "Base Compositional Bias and Phylogenetic Analyses: A Test of the Flying DNA Hypothesis." *Molecular phylogenetics and Evolution.* 13(3): 408-416

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

[35] Z. H. Yang. 2006. *Computational Molecular Evolution*. Oxford University Press.

[36] D.J. Zwickl, D. M. Hillis. 2002. "Increased Taxon Sampling Greatly Reduces Phylogenetic Error." *Syst. Biol.* 51(4):588-598