Spring 5-12-2017

Using Statistical Techniques to Estimate Rooted Species Trees from Unrooted Gene Trees

Ayed Rheal Alanzi
University of New Mexico, Albuquerque

Ayed Rheal A. Alanzi

Candidate

Mathematics and Statistics

Department

This dissertation is approved, and it is acceptable in quality and form for publication:

Approved by the Dissertation Committee:

James Degnan , Chairperson

Gabriel Huerta

Erik Erhardt

Christopher Witt

Using Statistical Techniques to Estimate Rooted Species Trees from Unrooted Gene Trees

by

Ayed Rheal A. Alanzi

B.E, Aljouf University, 2006
M.S., Statistics, Malaya University, 2009
M.S., Mathematics, Southern Illinois University, 2014

DISSERTATION

Submitted in Partial Fulfillment of the

Requirements for the Degree of

Doctor of Philosophy
Statistics

The University of New Mexico

Albuquerque, New Mexico

May, 2017

Dedication

I dedicate this work primarily to Allah (SWT),
and secondly to my parents,
wife and children, and all of my brothers and sisters.

Acknowledgments

First and foremost, I am very thankful to Allah, the most gracious and the most merciful.

As for the my fellow humans, I would first like to thank Professor James Degnan for guiding me step by step through my dissertation journey. Without his help, I would not have had the understanding to envision such an immense task. His patient way of dealing with my oversights has made this stressful event a learning experience which I have enjoyed much more than I thought I would. I am grateful to Professor Gabriel Huerta. My time studying with him first gave me the tools to work confidently with Bayesian Statistics. Without this competence, many aspects of this dissertation would have been of a diminished quality. I am also thankful for his help in speeding up my study by allowing me to bypass certain prerequisite formalities. I would also like to express my gratitude to Professor Erik Erhardt. My time peer-teaching a class on advanced data analysis with him has enlightened me on the clearest ways to explain data to an audience. The steps he takes when walking students through a problem shows a great understanding of what students do and do not understand. Even though I have not had the pleasure of working with Professor Christopher Witt prior to beginning work on my dissertation, his input has been highly appreciated and has formed an integral part of the revision process, and I am thankful for this.

I would like to express my love and gratitude to my father, Rheal Alanzi, and mother, Zahyaa Banikhalid, for supporting me throughout my education. Without their devotion to my education and their attention to my desire to go beyond a single degree, I would not have had the resources or the encouragement to go as far as I have. I am also grateful for the great example set for me by my brothers, Khalid, Saleh, and Muteb, and my sisters, Jawaher, Khalidiyyah, Fadyah, Tarfah, Saleha, and Afaf. The fact that all of them are high achievers has made it practically impossible for me to settle for mediocrity. The family members to whom I would like to give the warmest thanks are my wife, Fatmah, and my children, Abdulsalam, Shahad, Shaden, and Abdulmalik, since their love for me on a daily basis, as we sit together and spend time doing family activities, is what keeps me motivated to continue working hard even when I feel tired. Finally, there are countless individuals whose support I would like to thank, but I do not have room for all of their names.

Using Statistical Techniques to Estimate Rooted Species Trees from Unrooted Gene Trees

by

Ayed Rheal A. Alanzi

ABSTRACT OF DISSERTATION

Submitted in Partial Fulfillment of the
Requirements for the Degree of

Doctor of Philosophy
Statistics

The University of New Mexico

Albuquerque, New Mexico

May, 2017

Using Statistical Techniques to Estimate Rooted Species Trees from Unrooted Gene Trees

by

Ayed Rheal A. Alanzi

B.E, Aljouf University, 2006

M.S., Statistics, Malaya University, 2009

M.S., Mathematics, Southern Illinois University, 2014

PH.D., Statistics, University of New Mexico, 2017

Abstract

Methods for inferring species trees from gene trees motivated by incomplete lineage sorting typically use either rooted gene trees to infer a rooted species tree, or use unrooted gene trees to infer an unrooted species tree, which is then typically rooted using one or more outgroups. Theoretically, however, it has been known since 2011 that it is possible to infer the root of the species tree directly from unrooted gene trees without assuming an outgroup. The present work is the first that we know of which attempts to infer the root of a species tree using unrooted gene trees as the input data and without assuming an outgroup. It is hoped that this approach will be useful in cases where an appropriate outgroup is difficult to find and gene trees do not follow a molecular clock. The method uses Approximate Bayesian Computation

(ABC), and could also be useful when there is prior information that makes a small number of root locations plausible in an unrooted species tree. This study also uses the MLE method to compute the maximum value of the correct tree, which also uses bootstrapping to support the MLE results. Also, this study makes a comparison between using rooted gene trees and unrooted gene trees, both with and without DNA sequences, for five and eight taxa. Finally, an original method developed in this work is applied in an empirical study to data from Xi et al. (2014), and which uses their hypothesis as part of the prior in the present study.

KEY WORDS: multispecies coalescent, outgroup, midpoint rooting, molecular clock, identifiability, sufficiency, MLE, bootstrapping, DNA sequences.

Contents

List of Figures

List of Figures

List of Figures

List of Tables

List of Tables

Chapter 1

Introduction

This chapter contains four sections, all of which briefly show the history of the field. The first section explains rooted gene trees and unrooted gene trees as they occur in phylogenetics. The second section discusses the history of Bayesian inference with phylogenetics and the development of Approximate Bayesian Computation (ABC). Section three deals with the maximum likelihood in this field. Finally, this chapter concludes with a statement of this research.

Felsenstein (1981) goes over how techniques involving maximum likelihood are applied for purposes of estimating how data for sequences of nucleic acid derives evolutionary trees. He then develops a method for estimating ML, which he distinguishes from the probability that the tree itself is correct, and he shows this method to be computable through certain available programs. He contrasts his method with other types of algorithms, which he demonstrates provide highly flawed results.

In addition to the work on gene and species trees, many studies have also been done to compute ML trees from DNA sequences (Huelsenbeck and Hillis, 1993; Kuhner and Felsenstein, 1994; Huelsenbeck, 1995; Rosenberg and Kumar, 2001; Ranwez and Gascuel, 2002). The programs used in these studies can recover the correct tree

data sets to simulate data more frequently than other methods could have shown. As noted above, programs dealing with DNA sequencing problems are of the same nature as those dealing with gene-tree and species-tree problems, since they involve working with the same sort of molecular clock considerations.

A branching-process model involving probabilistics derives both an evolutionary tree and a model outlining DNA sequence evolution and change along the tree. According to Felsenstein (1981), attempting a first-process model is impeded in a number of ways. First, it requires both a process for speciation and one for extinction. It also requires a process which is particularly difficult to model, involving the selection of species from available candidates. Given such impediments, Felsenstein opts out of using a probabilistic branching model for hypothesizing an evolutionary tree, which he prefers to treat as an "unknown entity" (p. 369).

1.1 Gene Trees and Species Trees Background

A phylogenetic tree is called a species tree when it shows speciation events. It depicts the gradually evolving relationships among the set of biological species that share common ancestors. Species trees can be inferred through the data collected from multiple genes. To find out the most recent common ancestors of a gene from multiple species, a sample of genes is taken. Several examples illustrate how the gene tree can differ from its species tree. Persistence of ancestral polymorphisms can lead to this kind of difference, which is also called deep coalescence (Maddison, 1997). Two gene lineages coalesce (going backward in time) when they have a single ancestral gene as in Figure 1.1. The cause of deep coalescence is that gene copies from different species might fail to coalesce in their most recent ancestral population. Effective population size and speciation time can affect coalescence time. In coalescent theory, species trees and gene trees are rooted phylogenetic trees (Wakeley, 2009),

whereas gene trees estimated from molecular sequences are typically unrooted trees (Felsenstein, 2004).

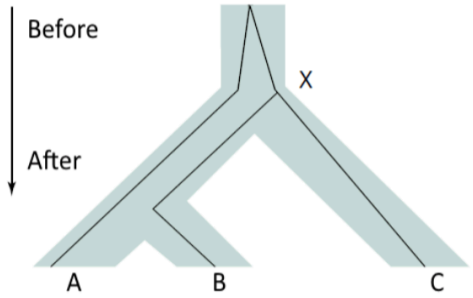

Figure 1.1: A species tree (outer-lines) with genes (inner-lines) coalescing back in time to a common ancestor gene (adapted from Nichols (2001), p. 359).

The Kingman coalescent model (Kingman, 1982) is a limiting case of the Wright-Fisher model (Wright, 1931) in which there is an infinite population and a reliance on continuous times. That leads to saying that under the Kingman model coalescence time follows the exponential distribution with the rate determined by the population size. The Kingman model is useful for determining the distribution of the gene trees for specific species trees (Pamilo and Nei, 1988; Degnan and Salter, 2005).

There are several software packages that can simulate gene trees within species trees such as SIMCOAL (Laval and Excoffier, 2004), ms (Hudson, 1983), and MaCS (Chen et al., 2009). Csilléry et al. (2010) give a full section in their article about the ABC software that researchers use for simulating gene trees. Fan and Kubatko (2011) also in their work used ms program to simulate the gene tree. According to Zhu et al. (2015), the program hybrid-lambda works to simulate the gene tree under Kingman's coalescent and other coalescent models. It deals with the specific sample size of species in the population of genes and allows for differentiation between gene

populations.

A lot of researchers like Nei (1987), Pamilo and Nei (1988), Rosenberg (2002), Degnan and Salter (2005), and Degnan and Rosenberg (2009) dedicate much of their research to the relation between species trees and gene trees by studying the origins of species through the rooted gene trees. In a similar attempt to examine gene tree history, Åkerborg et al. (2009), Liu et al. (2011), and Rasmussen and Kellis (2012) focus their work on birth and death processes. Three important studies require attention in order to fully understand the scope of research in unrooted gene-tree topology from the multispecies coalescent. Larget et al. (2010), Liu and Yu (2011), Mirarab et al. (2014), and Chifman and Kubatko (2014) focus on estimating the unrooted species trees from the topology of unrooted gene trees. Allman et al. (2011b) likewise studies the problem and reached a very important conclusion for unrooted gene tree topologies for five taxa or more, which is that it is possible to infer the rooted species tree topology by knowing the unrooted gene trees' topologies.

Rannala and Yang (2003) derive the density of rooted gene trees by working with the coalescence times and topologies. The methods described above use only topologies of gene trees. The characteristics of unrooted gene trees under coalescence still need development in order to fully understand the distribution of branch lengths.

Probabilistic model of DNA sequences, which evolve on gene trees, typically do not depend on the root location under most models because mutation is assumed to be reversible (Felsenstein, 1981). These considerations lead to phylogenetic programs which return unrooted gene trees. One thing that impedes phylogenetic programs from accurately rooting gene trees is the possible lack of an out-group or even of a molecular clock. As such, they are forced to rely on DNA sequence data alone, which produces limited results (Huelsenbeck et al., 2002; Boykin et al., 2010). A relaxed molecular clock uses Bayesian statistics and the maximum likelihood methods to provide a compromise between the molecular clock and the many-rates model, using

MCMC techniques to determine the parameters (Felsenstein, 2001). Even with rates of variation inside of lineages, such approaches can give accurate time estimates when a strict molecular clock is not applicable (Drummond et al., 2006).

In their analysis of the phylogenetic isolate Orcuttieae, Boykin et al. (2010) note that, while the outgroup method is most often used for phylogenetic tree rooting, it has a number of drawbacks and, as such, must be substituted with other methods, such as midpoint rooting, when data for an outgroup cannot be identified or is otherwise not available. To perform midpoint rooting, it is first necessary to estimate the ML tree. It is then necessary for the root to be placed on middle point of the ML tree's longest branch, assuming that the terminal at taxa at both ends have evolved at the same rate (Boykin et al. (2010), p. 688). This assumption, paired with the smaller number of taxa factored in, makes for a somewhat weak and at times unreliable analysis (Boykin et al. (2010), p. 688).

In the past, researchers assumed that gene trees had to be rooted under the coalescence model (Huang et al., 2010). However, the unrooted topological gene tree could be considered the occasion when one of its rooted types happens (Heled and Drummond, 2010). The formula $(2n-5)!!$ describes the number of unrooted gene trees from n species. There are $2n-3$ possible edges where the root can be located. According to Degnan (2013) as well as Degnan and Rosenberg (2006) the gene tree topology and the species tree topology may be occasionally different from each other (Degnan and Rosenberg, 2009). The authors refer to this condition as "the anomaly zone" if the most likely gene tree has a different topology than the species tree. In fact, there are no anomalies for the rooted species trees of three taxa or for unrooted species trees of four taxa (Degnan and Rosenberg, 2009; Allman et al., 2011b; Degnan, 2013). The probabilities of the unrooted gene tree can be found by the sum of the probabilities of the rooted gene trees with the same unrooted topology. The random variable of rooted gene trees follows a distribution which depends on

the species trees under the multispecies coalescent model. Figure 1.2 and Figure 1.3 provide graphic illustrations of the following formula: $P(U_1)$=$P(T_1)$+$P(T_2)$+ $P(T_3)$+$P(T_4)$+$P(T_5)$+$P(T_6)$ +$P(T_7)$.

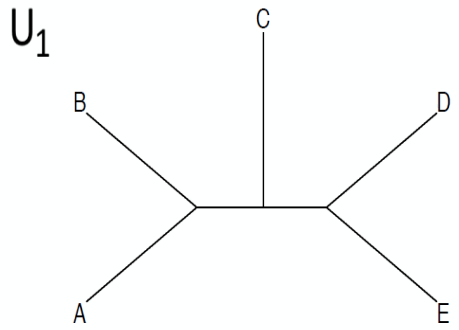

Figure 1.2: The unrooted gene trees topologies for 5-taxa species

By *nontrivial splits*, the topology of unrooted gene trees divides the taxa into two sets, which happens by removing one interior edge of the unrooted tree. Accordingly, "A set of all taxa descended from a node in a rooted tree forms a clade, the rooted analog of a split" (Allman et al. (2011b), p. 838). When separated, the interior branch or interior node of taxa in the phylogenetic trees connecting to two connected parts shows bipartition or split (Salichos et al., 2014). Both Semple and Steel (2003) and Chifman and Kubatko (2014) define a split in phylogenetics as the division of two exhaustive subsets of taxa that do not share any set members. All these definitions explain how the idea of the split would occur and also clarifies that the intersection between the two splits are empty because each one has the specific property that the other split does not have. For five taxa, it is evident that there are two taxa together on one side and the rest on the other side like $\mathcal{A}B|\mathcal{C}DE$, $\mathcal{A}C|\mathcal{B}DE$, which is clearer in reference to Table 2.1, taken from the Allman et al. (2011b) paper. It is also evident that there is no difference between the splits $\mathcal{A}|\mathcal{B}$ and $\mathcal{B}|\mathcal{A}$ (Semple and Steel, 2003).

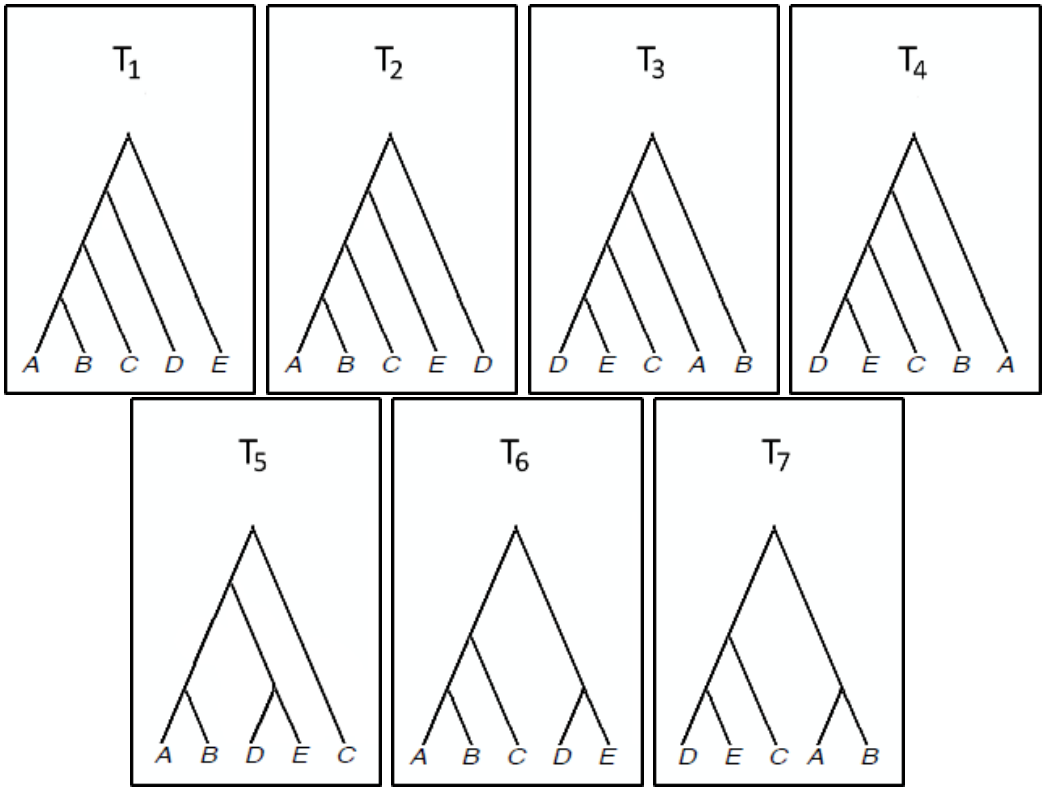

Figure 1.3: All shapes of rooted gene trees topologies for five taxa with the same unrooted topology as U_1 in Figure 1.2.

Allman et al. (2011b) notes that, if the probabilities of the unrooted gene topologies are known, then it is possible to get the topology of the rooted species tree. Figure 1.2 shows the topology of an unrooted gene tree for five taxa, which, according to Allman et al. (2011b), can be used to determine the topology of the rooted species trees. Figure 1.3 shows all seven of the possible rooted trees obtained by rooting the tree in Figure 1.2. Figure 1.4 similarly clarifies how it is possible to distinguish rooted gene trees from unrooted gene trees.

Box 1 of Figure 1.4 displays how the rooted gene tree appears when the root occurs on branch E of the unrooted gene tree. Then the unrooted gene tree topology gives the first rooted gene topology. In box 2 the point on the branch of D also leads to

the second rooted gene topology from the unrooted gene-tree topology. Box 3 shows the third shape of the rooted gene topology from the unrooted gene-tree topology. Box 4 presents the last *caterpillar* shape derived from the unrooted gene topology, when a point appears to Branch A, thus giving a rooted gene topology. A rooted *caterpillar* topology can be defined as a resolved tree in which every interior node has at least one leaf as its immediate descendant. In the topology of the unrooted gene tree in box 5, a point appears to Branch C, thus giving a rooted gene-tree topology, which is a unique shape called a *pseudocaterpillar*. In box 6, when A, B, and C coalesce with each other, the point in the branch between them and the coalescence of D and E gets the shape of species tree 6, called *balanced*. Finally, D, E, and C coalesce, and A and B coalesce. Then a point appears in the branch between them, thus giving a rooted gene-tree topology based on the unrooted gene-tree topology.

According to Drovandi and Pettitt (2012) and Robert (2016), ABC methods in statistics have become increasingly popular across the scientific fields in the last fifteen years, from epidemiology (Blum and Tran, 2010), to biology (Drovandi and Pettitt, 2011), to population genetics (Beaumont et al., 2002), since such methods proceed from model to data (inverting the more traditionally intuitive data to model), thus allowing statisticians to produce models that are increasingly more realistic and can handle evaluations that were previously deemed too computationally expensive, since they can now handle models that are capable of dealing with intractable likelihoods, for example, those which "cannot easily be completed or demarginalized by the introduction of latent or auxiliary variables" and which "cannot be estimated by an unbiased estimator" (Robert, 2016, p. 185).

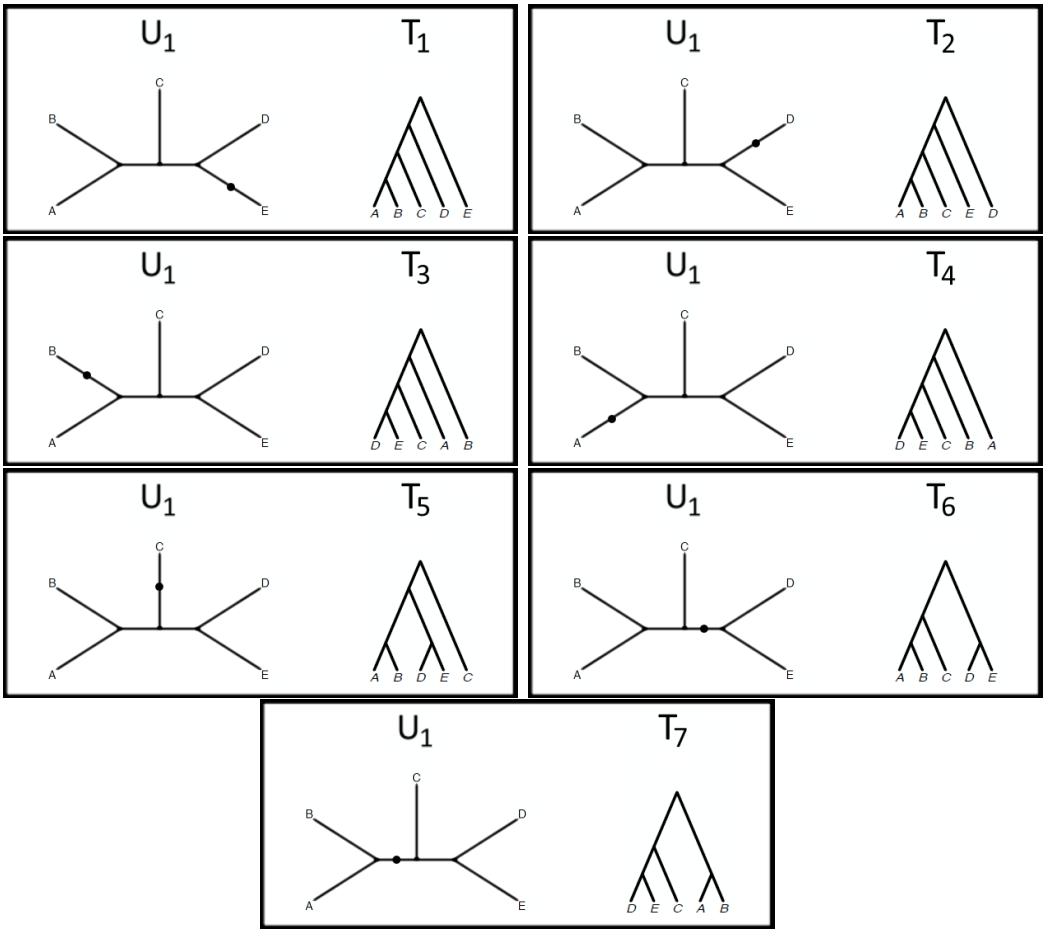

Figure 1.4: Unrooted gene-tree topology verses rooted gene-tree topology for 5 taxa

1.2 Approximate Bayesian Computation (ABC) Background

Bayesian approaches have a strong relationship with and are very similar to likelihood approaches. The main difference between the two approaches is that the Bayesian methods use the prior distribution but the likelihood methods do not deal with the prior distribution. In phylogenetics, there is a prior for the tree. According to Felsenstein (2004), Gomberg (1966) approved of the Bayesian method toward

inferring phylogenies using traits of species, which are modeled using a Brownian motion process. An important note for this paper is that it was an influential but not a published manuscript. The first appearance of the Bayesian methods in the phylogenetics field was in 1970.

Edwards (1970) was the first one who published the Bayesian approaches in this field, and he argued in favor of the ability of expecting the random patterns of branching and extinction on the tree, but that was not practical mathematically. Farris (1973,9) published two papers to support his discussion about the validity of the parsimony method by focusing on building basically a Bayesian model. Harper (1979) computed the probabilities of groups of species by using the Bayesian method, which takes information from back in the history of a clade. A clade can be defined as "A group of organisms that comprises the last common ancestor of these organisms and all extant and extinct descendants of that ancestor is called clade" (Sues (2016), p. 56). Wheeler (1991) discussed the correct evaluation for different data of phylogenies, which is taken from specific trees under the general parsimony. According to his assumption any step of using parsimony methods leads to the likelihood to decrease the factor of e. Smouse and Li (1987) similarly make the prior claim in the place of the individual three-tree topology and also calculate the posterior probabilities by the likelihood function also on the same three-tree topology; all this work was done with three taxa for rooted trees but they were not placed on the prior happening during the periods of the inside nodes for the tree. In its place, they also maximized the likelihoods to complete these and appropriated the outcomes for example of the likelihoods aimed at the tree topologies.

Most researchers who had discussed previously found it difficult to use Bayesian inference in full in phylogenetics. After all those researchers, Rannala and Yang (1996) got involved with their attempt to study the fully Bayesian model where they proposed the prior on the trees to be based on the birth-and-death process and to

analyze the DNA sequences by assuming the molecular clock.

Studies using Bayesian statistics in the field of phylogeny have led to a movement to further study the method of Approximate Bayesian Computation (ABC). A number of evolutionary biologists summarize its foundations and basic applications (Beaumont et al., 2010; Csilléry et al., 2010; Bertorelle et al., 2010). Woodhams et al. (2016) in their study give more elaborate examples about how to apply ABC to evolutionary biology. The examples are produced in two different types of simulations. The first type is the number of hybrid speciation events. The other one is the coalescence rate. They give an example using eight species of yeast, which come from a 106-gene data set from Rokas et al. (2003). 50% of the genome is taken from one parent; the simulation work in eight taxa and 106 gene trees allows for hybrid speciation events alone. The simulation is based on the coalescence rates and the hybrid speciation. They choose the iteration of simulation 100,000 after initial analyses. They used the method of ABC on the data of Rokas et al. (2003) according to the proposal of Fearnhead and Prangle (2012). Accordingly, the independent variable is the data set of simulation; the response variable is the logarithm of the coalescence rate. This is the way to produce the response variable, which is the log coalescence rate. This is also the same summary statistic that gave the hybrid speciation number of prediction. The result of the simulation parameters for ABC in Woodhams et al. (2016) conform to the output in Rokas et al. (2003), which consists of the yeast data set in 106 gene trees and eight species.

In population genetics, most regularly-used models include an enormous number of nuisance parameters which can be estimated through Bayesian methods (Shoemaker et al., 1999). BEST (Bayesian Estimation of Species Trees (Liu and Pearl, 2007)) and *BEAST (Bayesian Evolutionary Analysis Sampling Trees (Heled and Drummond, 2010)) are well-known approaches in species tree estimation. Using the coalescent model as the prior, BEST uses the program MrBayes to estimate the joint

posterior distributions of gene trees and species trees using Markov chain Monte Carlo (MCMC) for species trees as does *BEAST but with slightly different priors. Both BEST and *BEAST can use multilocus DNA sequence data as input but both are time consuming. There are other time efficient approaches such as STAR (species tree estimation using average ranks of coalescences (Liu et al., 2009)) and STEAC (species tree estimation using average coalescence times (Liu et al., 2009)). Approximate Bayesian Computation (ABC) is another good approach (Fan and Kubatko, 2011). The procedure of ABC is to simulate the data sets several times by using the prior distribution and then sufficient statistics or summary statistics are computed. If the distance is very close between the simulated sufficient statistics and the observed sufficient statistics, then the simulated parameters are accepted.

In addition, ABC is used as the second method with the Bayesian model and the summary statistic method. The rejection-sampling model for simulating a posterior for a parameter Φ is explained by Tavaré et al. (1997), which makes them the first researchers who introduced the ABC method. The ABC method assumes a prior distribution for Φ. Moreover, they used some summary, S, for the probability of observing the data be calculated and Φ' is drawn from the prior for Φ. After that, work is done on the rule of how to accept Φ' that has to work under this rule $P(S{=}s|\Phi') > cU$ with care given to the fact that U is the uniform distribution and c is the constant, which satisfies the condition $c < \max P(S = s|\Phi)$. These steps are reiterated several times, and the use of values acceptable for Φ' in order to form an estimate of the posterior distribution of Φ. Besides that, this approach is limited, since under this simple setup, it needs $P(\mathrm{s} = \mathrm{s}\ |\Phi)$ to be easily calculated and the maximum range to be Φ. Fu and Li (1997) develop this method by changing the accepted criterion $P(\mathrm{S{=}s}|\Phi') > cU$ by pairing the observed statistic and the values, which calculate the posterior probability for all of the data set. In addition, Weiss and von Haeseler (1998) disseminate this model to several summary statistics and simulate Φ' values using a network, so it has become the accepted norm $\|s' - s\| \leq \delta$,

where both of the s' and the s counts as vectors, tolerance δ and for several metrics $\|.\|$. Pritchard et al. (1999) corrected the problems with the model of Weiss and von Haeseler by simulating Φ from the prior distribution. But it has the restriction, which is that only a small number of summary statistics can be used. There are two developments produced from Beaumont et al. (2002) into the approach, which are the smooth weighting and the regression modification. This is the reason for raising the insensitivity from the approximate tolerance δ and allowing the use of more summary statistics. Moreover, the results become strong and accurate for summary statistics when used with the MCMC method, with which it is available to compare. Beaumont et al. (2002) studied the application of ABC to complex problems in population genetics.

Jensen et al. (2008) used the ABC method as an example to find the rate of selective sweeps in the sequence data of populations of the fly called *Drosophila melanogaster*. They also infer the relation of the parameters in regression, depending on the ABC model, and they find the sweep rate by using forward simulations in Drosophila melanogaster data. Csilléry et al. (2010) note that, even though the application of ABC models is itself easy, the process of making inferences still requires time-consuming steps which vary from situation to situation. Since data sets are often small or elliptical, tests must always keep in mind the potentially large margins of error. Buzbas and Rosenberg (2015) propose the idea of the Approximate Approximate Bayesian Computation (AABC). This is a class of mathematically possible ways to extend the ABC model spaces and provides an alternative to ABC in cases when the sample size is too small for ABC to process it. Since the margin of error for AABC is greater than it is for ABC, it should not be considered an alternative in cases when ABC can be used. ABC and AABC both use a mechanistic model for inferring model-specific parameters. However, ABC uses likelihood only derived from that mechanistic model, while AABC uses likelihood derived from a non-mechanistic model that simulates data from the mechanistic model.

Fan and Kubatko (2011) propose an ABC-inspired algorithm, which they call ST-ABC to estimate rooted species trees based on previous knowledge of rooted gene trees and to estimate the branch lengths of those rooted gene trees. The procedure of ABC is to simulate the data sets several times by using the prior distribution. After that, sufficient statistics are computed. If the distance is very close between the simulated sufficient statistics and the observed sufficient statistics, then the simulated parameters are accepted. They did the simulation by using the COAL program developed by Degnan and Salter (2005). Symmetric and asymmetric species trees served as subjects of the two types of simulations, and various branch lengths were taken into account, as were a number of sample sizes. The two sources of empirical data were yeast genes from seven taxa and primate genes from four taxa. According to their research, the posterior distribution of the parameters is estimated by the accepted values from the ABC method.

According to Buzbas (2012), they did not apply the ABC method correctly, and as such, it is not a reliable representative of ABC statistics. Since a proper ABC algorithm follows a different set of steps than Fan and Kubatko's methodology does, their proposed ST-ABC might not accurately produce a distribution which is similar to genuine posterior distributions or which reveals the actual species tree. This is because they compute the probability of the data under species trees sampled from the prior rather than simulating data sets. Buzbas (2012) furthermore claims that the work of Fan and Kubatko (2011) is not a reliable representative of ABC statistics because of its inability to correctly sample the posterior distribution; Buzbas develops this claim in his theoretical explanation which draws on three-taxon examples. A proper ABC algorithm follows a different set of steps than Fan and Kubatko's methodology does.

Kubatko and Fan (2013) respond by acknowledging that Buzbas was partly correct in his criticism because they use a computed distribution to make predictions

about how much the observed data should resemble the data simulated from the prior distribution. However, they emphasize that their algorithm (Fan and Kubatko, 2011) is not identical to that of a genuine ABC and that it is unreasonable to expect it to perform in the same way, as it is merely meant to perform a similar function when ABC is not available for application.

Nonetheless, Buzbas forces them to consider refinements to their theory, such as proposing the ST-ABC-CORRECTED algorithm. This algorithm requires continually estimating gene trees for sampled data sets. However, they ultimately do not recommend using such an algorithm because it would negate the efficiency gain of their original ST-ABC algorithm, and as such, they do not consider it useful.

An ABC method can deal with data simulated by models with noisy parameters. The combination of summary statistics and Bayesian statistics provides the advantages of computational convenience with resolving multiple parameter problems.

1.3 Maximum Likelihood Estimate (MLE) Background

Felsenstein (2004) recaps the history of likelihood methods starting with Edwards (1964), who introduced likelihood methods dealing with phylogenies for purposes of understanding the data of gene frequency. Felsenstein then discuss how Neyman (1971), who was originally critical of the use of likelihood methods, was the first to apply them to molecular sequences. Felsensten concludes by mentioning those who built on this, including the work of Kashyap and Subas (1974), as well as his own work, which applies to nucleotide sequences.

Knowles and Kubatko (2011) call attention to the fact that estimates for ML

species trees can be obtained in two successive stages. In the first of these stages, when working with a multi-locus data set for any given species tree, the evaluation of the likelihood function must be applicable. In the second stage, which relies on the success of the first, it is necessary to develop a method for locating a likelihood maximizing tree by combing through the array of possible species trees within a given space.

They go on to discuss the application of a likelihood function to both species trees and gene trees. The likelihood of species trees is computed in a variety of ways depending on whether the sample merely involves gene tree topologies, a gene tree topology with branch lengths, or multi-locus sequence data, whereas the likelihood of gene-tree data involves gene-specific DNA sequence data.

Two character-state methods which are of particular usefulness are the maximum parsimony (MP) method (Eck and Dayhoff, 1966) and the maximum likelihood (ML) method (Felsenstein, 1981). According to the parsimony method (Eck and Dayhoff, 1966), trees derived from fewer changes are more probable. As such, the method prefers to hypothesize the smallest number of mutations that can be used to account for any evolutionary change. According to the maximum likelihood method (Felsenstein, 1981), preference is not given to the smallest number of mutations; instead, preference is given to the mutations with the highest probability of working together to produce the observed data, according to stochastic models of nucleotide sequences. The best maximum likelihood estimate (MLE) for a tree can be determined by choosing the highest value of the MLE, which usually happens by estimating the ML of the branch lengths for specific tree topology and DNA substitution model. This process is done many times with other topologies (Felsenstein, 1981). The MLE method has an established statistical foundation (Felsenstein, 1981; Goldman, 1990) and is strong at restoring the true topology of trees by using a computer simulation study (Fukami-Kobayashi and Tateno, 1991; Hasegawa et al.,

1991). Bouckaert et al. (2014); Ronquist et al. (2012), and Alfaro et al. (2003) use the frequentist parametric approach to estimate tree phylogeny. They try to find the substitutes of the Bayesian and nonparametric approaches to estimate trees. The argument they make about its efficacy is part of a long tradition in the biology literature. Rogers (1997) and Yang (1994) developed the ideal characteristics of the MLE of asymptotic properties. The problem that they encountered was consistency. Yang (1994) has developed a guide to assess the consistency of ML trees that reflect the complexity of the problem.

Another important feature of the ML method is that it can calculate the different models of evolutionary trees in a statistical framework. In the last five years, many notable programs have been developed to also infer species trees. Wu (2012) introduces a new algorithm for species tree inference based on maximum likelihood. The likelihood is based on probabilities of rooted gene tree topologies. He compares his algorithm to the algorithm of Degnan and Salter (2005), and finds his algorithm is faster. He called his algorithm STELLS (which stands for Species Tree InfErence with Likelihood for Lineage Sorting). The methods used by Yu et al. (2011) and Yu et al. (2013) rely on hybridization as well as incomplete lineage sorting (ILS). Yu et al. (2013) try to solve the inference problem by exploring the space of phylogenetic networks, which they accomplish by using search heuristics in the software PhyloNet (Than et al., 2008). PhyloNet can infer species trees and networks using probabilities of rooted gene tree topologies. Zhu (2012) discusses the developments of Hybrid COAL, used for computing the probabilities of gene trees within a network. He suggests looking at probabilistic modeling coalescence with sorting in the lineages of hybrid species. In this work, trees always represent the relationships at the genetic level, and so he represents the relationships between species through a network instead of a tree.

1.4 Assumptions

The present work employs ABC to estimate the rooted species tree topology while pointing out that ST-ABC does not fit correctly in the Fan and Kubatko (2011) paper about ABC, as suggested by Buzbas (2012), who claims that their ST-ABC is not a reliable representative of ABC statistics because it might not accurately produce a distribution which is similar to genuine posterior distributions or which reveals the actual species tree. The present work uses simulation data with various parameters and sample sizes to investigate the performance of the ABC approach.

The present work also extends the rest of the equations that Allman et al. (2011b) mention in their article for five species, and both calculates them, and then uses statistical tools, such as maximum likelihood and bootstrapping, for inference. The present work, moreover, uses equations to compute the maximum likelihood estimate (MLE); to calculate this, the present work uses the bootstrap analysis to estimate support for the maximum likelihood species trees.

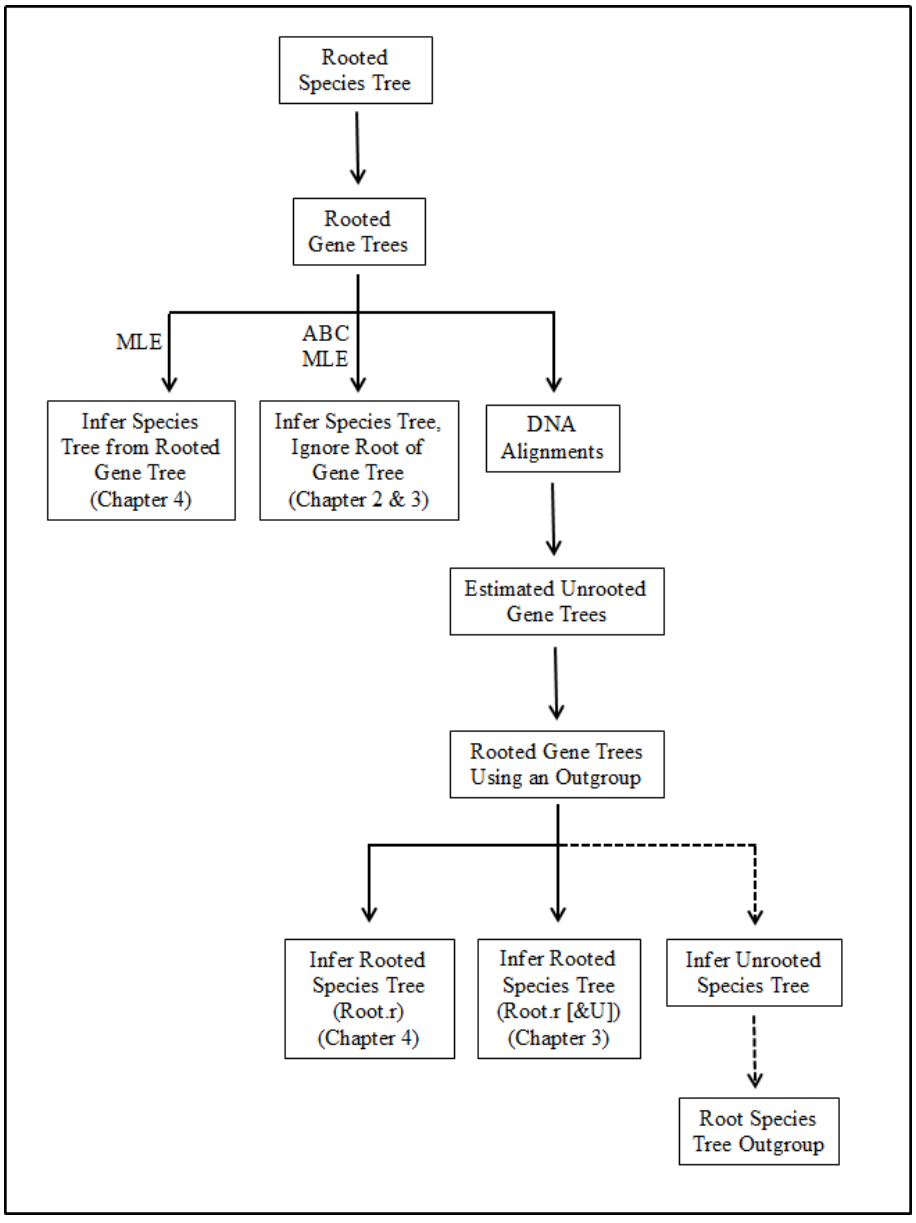

Figure 1.5: Tree of all Chapters in the Dissertation.

Chapter 2

Approximation Bayesian Computation (ABC)

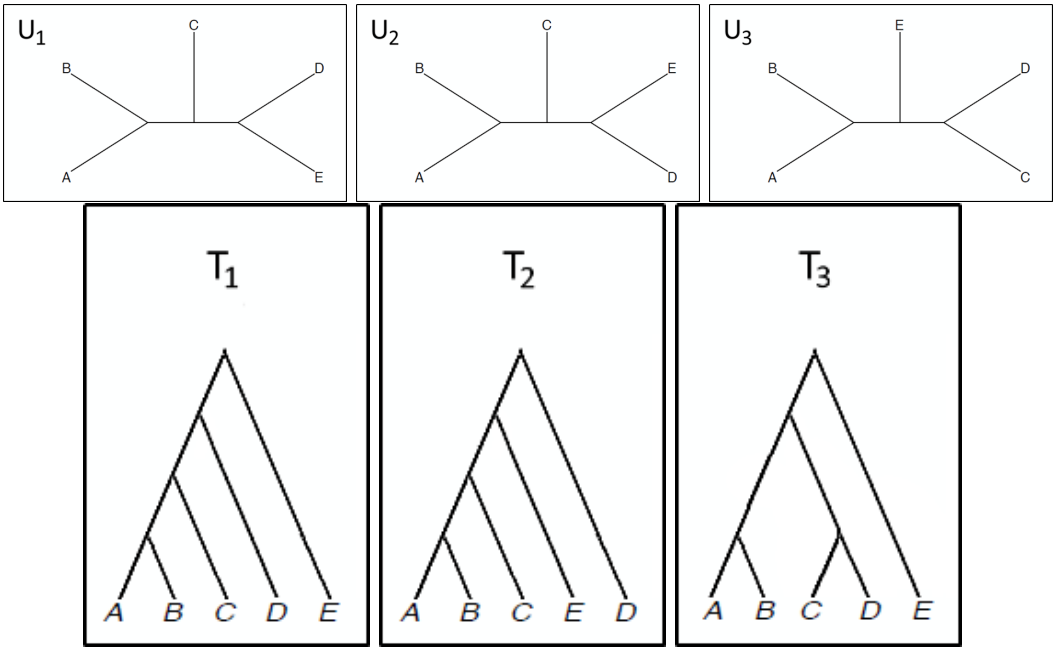

Figure 2.1: Example of counting topologies for 5-taxa species. Topologies T_1, T_2, and T_3 all have the topology U_1 when they are unrooted.

This chapter explains the ABC method and the simulation study. It also shows the results of this method for five and for eight taxa. Before starting with explanations of the method and the simulation, it is necessary to explain the way to calculate the topology and the split frequencies. An example is given in Figure 2.1, which is the first way to count the topology for unrooted gene trees and rooted gene trees. The "n.topo.obs" is defined as a vector where the i^{th} entry is the number of times topology U_i is observed. For example, the number of topologies observed in "n.topo.obs" is (2, 0, 1, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0), where the first two unrooted gene tree topologies are presented in the same way as the two rooted gene tree topologies, and the third unrooted topology is presented as equal to the two rooted gene tree topologies. The "n.split.obs" is defined as a vector where the i^{th} entry is the i^{th} split based on the 10 splits listed in Table 2.1. The count of the splits observed in "n.split.obs" is calculated as (3, 2, 0, 0, 1, 0, 0, 0, 0, 0), and the 3 comes from observing the split AB|CDE three times and ABC|DE twice. Lastly the CD|ABE split occurs once.

2.1 Method

The general methodology for ABC is to first simulate from the prior distribution for the parameter (in this case, a species tree with branch lengths), then to simulate data from parameter (gene trees from species trees), and to record a distance between the simulated data set and the observed data. In this project, the observed data as well as simulated data consist of unrooted gene tree topologies. The distance computed between the observed data and the simulated data depends on a choice of summary statistic, and a few variations were used. For 5 taxa, there are only 15 unrooted gene tree topologies, and it is possible to record how often each of the 15 gene tree topologies occurs for both the observed and simulated data. A vector of topology

counts therefore is a summary of the data which preserves all relevant information in the sample. A variation is to instead record the splits that occur in a sample of trees, counting how often each split occurs, but discarding information about which split came from which tree. For five taxa, there are 10 possible splits.

For 8 taxa, there are 10,395 possible unrooted tree topologies. Having a vector of length 10,395 to record counts of how often each possible topology occurs is impractical, and is an example of the difficulty of using ABC in high-dimensional problems. Fan and Kubatko (2011) deal with this problem for 8 taxa by only recording counts of the rooted gene trees occurring in the observed data, and then counting the number of gene tree topologies in the simulated data that correspond to one of the input trees. However, the number of distinct tree topologies can still be quite high (Salichos and Rokas, 2013), and Fan and Kubatko (2011) found their approach to be less accurate for trees with eight tips than for four, which might speculatively be due to the dimensionality problem. consequently for 8 taxa, only splits are used, and instead of recording counts of all possible splits, the symmetric difference between the set of splits in the observed data versus the simulated data is record. Details of these algorithms are given below.

The data is composed of counts of L unrooted topologies for 5 taxa presented in the observed data. The present work records "n.topo.obs", which is the number of times each unrooted gene tree topology was observed. The ABC algorithm for inferring the rooted topology of the species tree involves steps 2-7. Step 1 is used to simulate the observed data. Let G_n denote the number of unrooted topologies for n taxa.

Algorithm 1

1. Simulate the observed data by using the program Hybrid-Lambda (Zhu et al., 2015) to simulate from the species tree, which is called "n.topo.obs".

2. Start with $j = 1$.

3. Simulate a species tree from the prior distribution of rooted species trees.
4. Simulate gene trees from the species tree sampled from the prior by using the program Hybrid-Lambda; the vector of gene tree counts is called "*n.topo.sim*".
5. Calculate D_j.topo $= \sum_{i=1}^{G_n}(n.topo.obs_i - n.topo.sim_i)^2$
6. Increment j by 1 and repeat steps (2) – (5) J times.
7. Take the smallest αJ values from step 5, then retain the species trees corresponding to these smallest distances. These trees estimate the posterior distribution.

By finding a way to summarize the retained species trees it is possible to estimate their topology. For example, in this study the α= 0.002 and $J = 50,000$ were used to get the 100 trees with the smallest distances.

The present work infers the rooted topology of the species tree by summarizing the posterior distribution of the species tree topologies. For species trees with a large number of taxa, it is possible to use a consensus tree for the topology estimate.

The same method that was implemented for the topology counts can instead use split frequencies of the gene trees. The data is summarized by the counts of splits in L unrooted gene trees for 5 taxa presented in the observed data, which is recorded as "n.split.obs", and this is the number of times that each split was observed. We let S_n denote the length of the vector "n.split.obs" and the formula for S_n:

$$S_n = \begin{cases} \sum_{i=1}^{\frac{n}{2}-1} \binom{n}{i} + \frac{1}{2}\binom{n}{\frac{n}{2}} & \text{if } n \text{ is even} \\ \sum_{i=1}^{\frac{n-1}{2}} \binom{n}{i} & \text{if } n \text{ is odd} \end{cases}$$

The ABC algorithm, depending on the split counts, involves the following steps:

Algorithm 2

1. Simulate the observed data by using the program Hybrid-Lambda to simulate gene trees, and count the splits in the gene trees which is called "n.split.obs".

2. Start with $j = 1$.

3. Simulate a species tree from the prior distribution of rooted species trees.

4. Simulate gene trees from the species trees sampled from the prior distribution by using the program Hybrid-Lambda. The count of the splits is called "$n.split.sim$".

5. Calculate D_j.split $= \sum_{i=1}^{S_n}(n.split.obs_i - n.split.sim_i)^2$

6. Increment j by 1 and repeat steps (2) – (5) J times.

7. Take the smallest α J values from step 5, then retain the sample species trees corresponding to these smallest distances. Those trees estimate the posterior distribution.

The above two algorithms were applied for trees with five taxa. For trees with eight taxa, splits are recorded as multisets, which are sets that keep track of the multiplicity of each element—the number of times each element of the set occurs, involves the following steps:

Algorithm 3

1. Extract splits from the observed gene trees in to the multiset S_{obs}.

2. Start with $j = 1$.

3. Simulate a species tree from the prior distribution of rooted species trees.

4. Simulate gene trees from the species trees sampled from the prior distribution by using the program Hybrid-Lambda. Extract all splits from the simulated gene trees into the multiset S_{sim}.

5. Let $D = |S_{obs} \setminus S_{sim}| + |S_{sim} \setminus S_{obs}|$

6. Increment j by 1 and repeat steps (2) $-$ (5) J times.

7. Take the smallest αJ values from step 5, then retain the sample species trees corresponding to these smallest distances. Those trees estimate the posterior distribution.

In these implementations of the ABC approach, the best αJ species trees are retained to estimate the posterior distribution. In original descriptions of the ABC

algorithm, the idea was to accept simulated parameters when the distance satisfied $D < \delta$ where δ was some tolerance. This approach leads to a random number of parameters accepted in forming the posterior distribution. The alternative method of retaining the best αJ trees corresponds to setting δ to be some quantile of the observed distribution of D. This approach is used, for example, by Beaumont et al. (2002) and Nunes et al. (2010). The latter authors give an example with $J = 10^6$ and $\alpha = .01$, leading to accepting 10^4 sampled parameters. We used $J = 5 \times 10^4$ and $\alpha = .002$ to retain the best 100 trees.

The present work summarizes the posterior distribution of the species tree splits to infer the rooted splits. The scripts and R code applied can be seen in their entirety in Appendix A. To implement all calculations, this study employs a number of scripts, including the R-Package (Ihaka and Gentleman, 1996). These can be seen in their entirety in Appendix A.1, which displays the coding of the 5 taxa that are used in this study. Moreover, Appendix A.2 displays the code that calculates the 5 taxa from the DNA. Finally, Appendix A.3 also shows the code used for calculating the 8 taxa. All this script and code is used for calculating the posterior probability and maximum likelihood estimation.

This study is similar to Fan and Kubatko (2011) in terms of estimating species trees based on steps (2) to (4). However, in their study, they compute expected counts of gene trees theoretically, whereas gene trees were simulated in order to use a genuine ABC algorithm. This is also motivated by the fact that determining the expected counts requires computing an entire gene tree distribution, which is computationally expensive for larger trees (Degnan and Salter, 2005; Wu, 2012). Second, the way that they computed the distance between the observed gene tree and the simulated gene tree used the formula $D_j = \sum_{i=1}^{G} \frac{(n_{obs,i} - n_{exp,i})^2}{n_{exp,i}}$, whereas the present study doesn't divide by the expected counts (see step (5) in algorithms 1–3. Third, they inferred the rooted species tree directly from rooted gene trees, but

in this study rooted species trees are estimated by using the unrooted gene trees. Finally, the priors used were restricted to locating the root of the species tree on an unrooted tree. In their case, the prior for the species tree was based on trees that were short distances from the input gene trees.

Table 2.1: The 15 unrooted topological gene trees for 5-taxa.

Tree	Splits	Probability
T_1	AB|CDE, ABC|DE	$u_1 = r_1 + r_2 + r_{59} + r_{60} + r_{67} + r_{76} + r_{105}$
T_2	AB|CDE, ABD|CE	$u_2 = r_3 + r_4 + r_{53} + r_{54} + r_{64} + r_{79} + r_{104}$
T_3	AB|CDE, ABE|CD	$u_3 = r_5 + r_6 + r_{47} + r_{48} + r_{61} + r_{88} + r_{103}$
T_4	AC|BDE, ABC|DE	$u_4 = r_7 + r_8 + r_{57} + r_{58} + r_{70} + r_{77} + r_{102}$
T_5	AC|BDE, ACD|BE	$u_5 = r_9 + r_{10} + r_{41} + r_{42} + r_{65} + r_{82} + r_{101}$
T_6	AC|BDE, ACE|BD	$u_6 = r_{11} + r_{12} + r_{35} + r_{36} + r_{62} + r_{91} + r_{100}$
T_7	AD|BCE, ABD|CE	$u_7 = r_{13} + r_{14} + r_{51} + r_{52} + r_{71} + r_{80} + r_{96}$
T_8	AD|BCE, ACD|BE	$u_8 = r_{15} + r_{16} + r_{39} + r_{40} + r_{68} + r_{83} + r_{95}$
T_9	AD|BCE, ADE|BC	$u_9 = r_{17} + r_{18} + r_{29} + r_{30} + r_{63} + r_{94} + r_{97}$
T_{10}	AE|BCD, ABE|CD	$u_{10} = r_{19} + r_{20} + r_{45} + r_{46} + r_{72} + r_{87} + r_{89}$
T_{11}	AE|BCD, ACE|BD	$u_{11} = r_{21} + r_{22} + r_{33} + r_{34} + r_{69} + r_{86} + r_{92}$
T_{12}	AE|BCD, ADE|BC	$u_{12} = r_{23} + r_{24} + r_{27} + r_{28} + r_{66} + r_{85} + r_{98}$
T_{13}	BC|ADE, ABC|DE	$u_{13} = r_{25} + r_{26} + r_{55} + r_{56} + r_{73} + r_{78} + r_{99}$
T_{14}	BD|ACE, ABD|CE	$u_{14} = r_{31} + r_{32} + r_{49} + r_{50} + r_{74} + r_{81} + r_{93}$
T_{15}	BE|ACD, ABE|CD	$u_{15} = r_{37} + r_{38} + r_{43} + r_{44} + r_{75} + r_{84} + r_{90}$

The present work uses Table 5 in the appendix of Allman et al. (2011b), as mentioned in Table 2.1. It is then necessary to take this table to follow it as a guideline. The first column in Table 2.1 is T_i where i is from 1 to 15, which indexes the unrooted topology for each tree. The second column in Table 2.1 is the split of each tree. The last column in Table 2.1 is the probability of each tree topology where r_i is the probability of the i^{th} rooted gene tree topology and u_i is the probability of the i^{th} unrooted gene tree topology (Allman et al., 2011b).

2.2 Simulation

A set of 14 5-taxon rooted bifurcating trees was selected by taking 5 rooted bifurcating tree shapes from the caterpillar tree, 4 rooted bifurcating tree shapes from the pseudocaterpillar tree, and 5 rooted bifurcating trees from the balanced tree to simulate the observation data. The species tree is the parameter, and the data consist of the gene trees. The three rooted bifurcating tree shapes are shown in Figure 2.2. The branch lengths used to simulate the observation data were (x, y, z) = (0.1,0.1,1.0), (1.0,1.0,1.0), and (0.1,0.1,1.0) (Fig. 2.2). After simulating the observation data, calculating the topology of the species, and recording them as observation topology data, the simulation is called "n.topo.obs". Furthermore, it is then necessary to calculate the splits of the gene trees, which is called "n.split.obs".

The second step is simulating data by using a prior. A uniform prior over the 7 trees with the same unrooted topology was used for the species tree topologies. An exponential distribution with rate one was used for branch lengths of the species tree. Once the data are simulated with those priors it is then necessary to compute the vector of topology frequencies, which is called "n.topo.sim". It is also necessary to calculate the split frequencies, which is called "n.split.sim".

All those simulations are done by using a Hybrid-Lambda program (Zhu et al., 2015). The simulation is done primarily with $J = 50,000$. For each repetition of J, a sample size of 100 genes was used. The value of α was set to 0.002 so that the 100 best species trees were retained. The formula αJ was used to determine the desired number of species trees based on the smallest D_j.

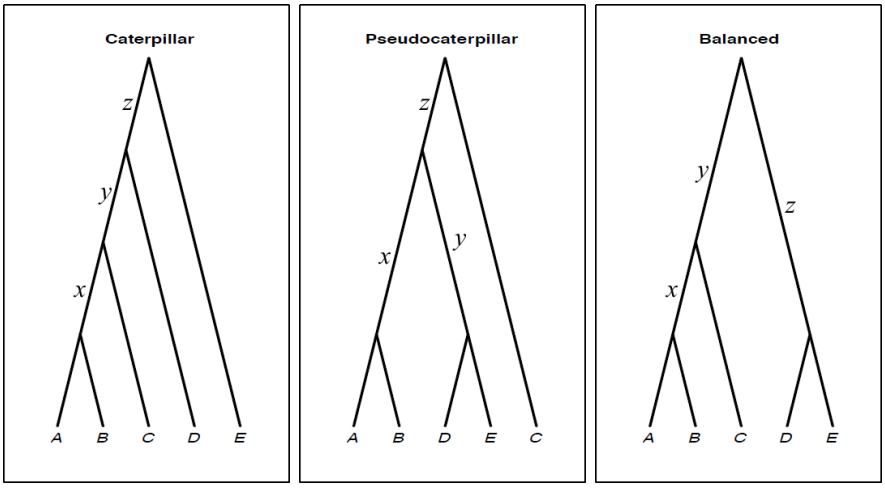

Figure 2.2: The rooted bifurcating tree shapes for 5-taxa species

2.3 Results

An important definition in this study which is used to summarize the results is as follows. The first thing is proportion correct, which refers to how many times the simulation study matches the correct tree out of all iterations. The second is the coverage probability, which refers to the percentage of times that the true tree is included in the 90% credible interval. The number of trees in the 90% credibility region (CR) refers to how many trees make the 90% CR even when this 90% CR does not include the correct tree.

Figure 2.3 shows all of the 5-taxon species trees used in the present work, including their branch lengths. Moreover, the first five species in Figure 2.3 have the caterpillar species tree shape, and the second five in the figure shows the balance species tree shape, and the last four species in the figure shows the pseudocaterpiller species tree shape. Figure 2.3 clearly shows that, when the branch lengths are different, the species trees look different.

To decide on a value of J, a pilot study was done with species tree 1. Table 2.2

shows all different numbers of J; the posterior probability using topology counts gives a better result than the posterior probability using split counts. And it is also evident that, when J=50,000 with 100 genes, it gives a higher posterior probability than for J=10,000 with 500 genes although the computation time is the same. Therefore, J= 50,000 was used in subsequent simulations.

Table 2.2: Posterior probabilities times 100 for the seven possible root locations. The posterior probability for the true tree is in bold.

J	No. of Gene	Posterior Probability of Topology Trees							Posterior Probability of Split Trees						
		1	2	3	4	5	6	7	1	2	3	4	5	6	7
50000	100	**67**	28	0	0	4	1	0	**45**	41	0	0	11	2	1
20000	100	**56**	32	0	0	6	3	3	**42**	34	3	1	12	4	4
10000	100	**52**	26	1	2	9	5	5	**36**	22	6	9	17	5	5
10000	500	**56**	40	0	0	0	3	1	**47**	45	0	1	1	4	2

Table 2.4 shows the result of the posterior probability of the topology of 5 taxa for all species trees used in the simulation. The best match among the caterpillar species trees is Species Tree 2, since it has the highest proportion correct at 90%, the highest coverage probability at 100%, the highest average posterior probability at 66.78% (for Posterior Species Tree 1), and the smallest credibility region at 2.4 trees. The second best match among the caterpillar species trees is Species Tree 1, and the other species trees are difficult to make inferences with, because they lack either a high enough proportion correct, a high enough coverage probability, a high enough average posterior probability, or a small enough credibility region.

The best match among the balanced species trees in Table 2.4 is Species Tree 7, since it has the highest proportion correct at 65%, the highest coverage probability at 92%, and the highest average posterior probability at 28.74% (for Posterior Species Tree 6); however, its credibility region of 5.26 is not the smallest among the balanced

Table 2.4: Posterior Topology for Five Taxa without DNA

Type of Tree	Species Tree (x, y, z)	Prop. Correct	Coverage Probability	Average of the Posterior Probability							# of trees in 90% CR.
				1	2	3	4	5	6	7	
Caterpillar	(0.1,0.1,0.1)	0.76	0.96	53.62	23.96	2.40	2.60	7.26	7.46	2.7	3.04
	(0.1,0.1,1.0)	0.94	1.00	67.66	20.8	0.30	1.64	4.04	4.7	0.86	2.4
	(1.0,0.1,0.1)	0.50	0.96	40.10	34	0.56	0.56	15.48	3.94	5.36	3.32
	(0.1,1.0,0.1)	0.47	1.00	36.06	33	0.28	0.36	1.3	28.86	0.14	3.06
	(1.0,1.0,1.0)	0.36	0.94	25.34	25.32	2.10	2.82	9.46	28.08	6.88	4.14
Balance	(0.1,0.1,0.1)	0.02	0.50	30.36	28.18	5.54	4.8	16.46	10.82	3.84	3.76
	(1.0,1.0,1.0)	0.65	0.92	15.94	14.82	8.52	8.08	13.02	28.74	10.88	5.26
	(0.1,0.1,1.0)	0.32	0.84	11.16	9.70	12.30	15.2	23.22	24.16	4.26	4.46
	(0.1,1.0,0.1)	0.26	0.88	34.6	36.06	0.54	0.38	1.28	27.02	0.12	2.98
	(1.0,0.1,0.1)	0.00	0.24	29.00	30.68	0.80	1.82	21.84	7.14	8.72	3.76
Pseudocater.	(0.1,0.1,0.1)	0.92	0.96	7.74	8.86	8.20	7.78	56.06	5.98	5.38	3.66
	(1.0,1.0,1.0)	0.09	0.54	16.32	16.18	12.12	13.78	12.46	14.96	14.18	4.40
	(0.1,0.1,1.0)	0.94	1.00	4.14	4.34	3.32	2.66	79.54	3.38	2.62	2.14
	(0.1,1.0,0.1)	0.52	0.96	2.46	2.84	24.22	20.90	31.74	12.80	5.02	4.22

species trees. The balanced species tree which is the most difficult to make inferences with is Species Tree 10, because the proportion correct is as low as possible at 0%, the coverage probability is very low at 24%, the average posterior probability is very low at 7.14%; however, its credibility region of 3.76 is not undesirably large among the balanced species trees.

The best match among the Pseudocaterpillar species trees in Table 2.4 is Species Tree 13, since it has the highest proportion of correct matches at 94%, it is

Table 2.5: Posterior Split for Five Taxa without DNA

Type of Tree	Species Tree (x, y, z)	Prop. Correct	Coverage Probability	Average of the Posterior Probability							# of trees in 90% CR.
				1	2	3	4	5	6	7	
Caterpillar	(0.1,0.1,0.1)	0.76	0.94	47.96	24.74	4.46	4.04	10.28	6.10	2.42	3.46
	(0.1,0.1,1.0)	0.94	1.00	59.6	27.64	0.96	2.00	5.20	3.64	0.96	2.72
	(1.0,0.1,0.1)	0.52	0.94	37.7	32.32	0.66	0.48	19.08	3.68	6.08	3.34
	(0.1,1.0,0.1)	0.35	1.00	35.30	33.28	0.80	0.46	2.16	27.80	0.20	3.06
	(1.0,1.0,1.0)	0.33	0.94	24.46	24.62	2.04	2.54	11.34	26.88	8.12	4.20
Balance	(0.1,0.1,0.1)	0.00	0.34	27.78	26.60	7.06	7.60	19.00	8.86	3.10	4.14
	(1.0,1.0,1.0)	0.72	0.92	15.76	15.26	9.08	8.08	12.06	28.84	10.92	5.40
	(0.1,0.1,1.0)	0.40	0.84	9.54	9.36	14.60	15.90	23.16	23.62	3.82	4.62
	(0.1,1.0,0.1)	0.36	0.94	33.68	33.22	1.10	1.14	2.12	28.58	0.16	3.16
	(1.0,0.1,0.1)	0.00	0.22	28.28	29.14	0.80	1.58	24.32	6.34	9.54	3.86
Pseudocater.	(0.1,0.1,0.1)	0.76	1.00	9.74	10.60	10.52	10.88	48.8	4.72	4.70	4.14
	(1.0,1.0,1.0)	0.10	0.56	16.34	15.60	11.94	13.38	13.04	15.84	13.84	4.52
	(0.1,0.1,1.0)	0.96	1.00	7.00	7.20	6.12	5.38	66.96	3.64	3.56	3.52
	(0.1,1.0,0.1)	0.56	0.94	2.44	2.52	24.30	23.24	30.54	12.38	4.50	4.22

included in all 90% of the coverage probability, and its posterior probability has an average of 79.54% among the rest of the Pseudo-caterpillar species trees, which is the smallest credibility region at an average of 2.14 trees. The second best match among the Pseudo-caterpillar species trees is Species Tree 11. The posterior probability of Species Tree 12 is the most difficult to infer since it has a 9% proportion of matches with the correct tree. Also the posterior probability of Species 12 has an average of 12.46, which is very small compared to the rest.

Table 2.5 shows the result of the posterior probability of the splits of 5 taxa

for all 5-taxon species trees in the simulation. The best match among the caterpillar species trees is Species Tree 2, since it has the highest proportion correct at 94%, the highest coverage probability at 100%, the highest average posterior probability at 59.12% (for the Posterior Species Tree 1), and the smallest credibility region at 2.7 trees. The second best match among the caterpillar species trees is Species Tree 1, and the other species trees are difficult to make inferences with, because they lack either a high enough proportion correct, a high enough coverage probability, a high enough average posterior probability, or a small enough credibility region.

The best match among the balanced species trees in Table 2.5 is Species Tree 7, since it has the highest proportion correct at 72%, the highest coverage probability at 92%, and the highest average posterior probability at 28.84% (for the Posterior Species Tree 6); however, its credibility region of 5.4 is not the lowest desirable among the balanced species trees. The balanced species trees which are the most difficult to make inferences with are Species Tree 6 and Species Tree 10, because the proportion correct for each one is as low as possible at 0%, the coverage probability for each one is very low at 34% and 22% respectively, the average posterior probability for each one is very low at 8.86 and 6.34 respectively; however, their credibility regions of 4.14 and 3.86 are not undesirably high among the balanced species trees.

The best match among the Pseudocaterpillar species trees in Table 2.5 is Species Tree 13, since it has the highest proportion of correct matches at 96%, the highest coverage probability at 100%, the highest average posterior probability at 66.96 (for the Posterior Species Tree 5), and the smallest credibility region at 3.52. The second best match among the Pseudocaterpillar species trees is Species Tree 11, and the other species trees are difficult to make inferences with, because they lack either a high enough proportion correct, a high enough coverage probability, a high enough average posterior probability, or a small enough credibility region.

Upon comparing the posterior probabilities based on topologies versus splits, it

is clear that the posterior probability based on splits gives a slightly lower probability on average than the posterior probability based on topologies for the highest posterior topology to match the rooted species tree. Also, the average number of trees in the 90% critical region in the posterior based on splits is more than it is for topologies, according to the results that were obtained from the simulation number $J = 50,000$ with 100 genes. The result of Species 2 from the caterpillar shape and Species Tree 13 from pseudo-caterpillar shape with the same branch length $(0.1, 0.1, 1.0)$ have a higher average posterior probability in topology and split cases; it is included in all 90% CR, and it also has a higher percentage of matches with the true tree. In two cases, the caterpillar trees and pseudocaterpillar trees agree in terms of branch-length results, which are $(0.1, 0.1, 0.1)$ and $(0.1, 0.1, 1.0)$. Among all the rest of the five species trees, Balanced Species Tree 10 with the branch length $(1.0, 0.1, 0.1)$ is the most difficult to infer.

Figure 2.4 shows the correlation between posterior probabilities based on topology counts vs. split counts of the topology matching the true tree and the posterior probability of splits for trees matching the true tree for caterpillar trees. Inside each chart of Figure 2.4 is the value of the correlation. Figure 2.4 shows everything from Species 1 to Species 5, which demonstrates that the true tree has to be Tree 1 since all five are caterpillar trees. According to this result, Species tree 1 has the higher correlation value, which means it shows the strongest relationship among trees. From among all the caterpillar trees, Species 4 has the lowest correlation between the posterior probability of the topology and the posterior probability of the split, but there is still a strong relationship between the posterior probability of topology for the true tree and the posterior probability of the split for the true tree.

Figure 2.5 shows the correlation between posterior probabilities of the topology matching the true tree using topology counts vs. split counts with balanced trees. Inside each chart of Figure 2.5 is the value of the correlation. Figure 2.5

shows everything from Species tree 6 to Species tree 10, which are all balanced trees. According to this result, Species tree 7 shows the strongest correlation between all of the balanced trees, and Species tree 6 shows the lowest correlation, but there is still a moderate relationship between the posterior probabilities based on topologies vs. splits. Therefore, Species Tree 7 has much evidence for the true species with this branch length (((A:1.0,B:1.0):1.0,C:2.0):1.0,(D:2.0,E:2.0):1.0).

The relationship between posterior probabilities of matching the true tree using topology counts vs. split counts with pseudocaterpillar trees is shown in Figure 2.6. Inside each chart of Figure 2.6 is the value of the correlation. All species trees, from Species Tree 11 to Species Tree 14, are shown in Figure 2.6, which are all Pseudocaterpillar trees. According to this result, all the Pseudocaterpillar trees have a stronger correlation with the posterior probabilities based on topologies counts vs. splits counts. Therefore, Species Tree 11 has much evidence for the true species with this branch length (0.1, 0.1, 0.1), which agrees with the highest correlation of caterpillar tree with the same branch length.

Figure 2.4, Figure 2.5, and Figure 2.6 demonstrate that Species Tree 1, Species Tree 7, and Species Tree 11 have the highest correlation between the topology and the split posterior probability matching the true tree. At the opposite end of the spectrum, Species Tree 6, from the balanced tree, has the lowest correlation among all species, which leads to the conclusion that Species Tree 6 is the most difficult to infer with these branch lengths (0.1, 0.1, 0.1). Moreover, Species Tree 11, from the Pseudocaterpillar tree, has the highest correlation when compared with all species together, which leads to the conclusion that it is the best species with this branch length (0.1, 0.1, 0.1), which agrees with the second highest correlation among all others species trees that also agree with Species Tree 1 from the caterpillar tree with the same branch length.

Figure 2.7 shows a correlation with the average posterior probability using

topology counts versus split counts for 5-taxon trees. The line $y = x$ is plotted so that points below the line indicate that the posterior probability was higher using topology counts rather than split counts as the summary statistic. Posterior probabilities for the correct tree tended to be slightly lower when using splits, but are highly correlated with posterior probabilities using topology counts – as shown in Figure 2.7 the correlation is larger than 0.99. There is a significant correlation between the posterior probability of topology and the posterior probability of a split in matching a correct tree in all of the species trees in this study.

2.4 DNA Sequences For Five species

The present study simulated DNA sequences by choosing the three 5-taxon species trees, which are one caterpillar tree, one balanced tree, and one pseudocaterpillar tree. Then, the present work applies the same ABC-algorithm (1 and 2) that is used to simulate the regular five taxa from the caterpillar species tree, the balanced species tree, and the pseudocaterpillar species tree. The present study uses the Seq-Gen program (Rambaut and Grassly, 1997) to generate DNA sequences of length 500 nucleotides using an $HKY + \Gamma$ model with base frequencies of $0.3, 0.2, 0.2$, and 0.3 for A, C, G, and T, respectively. The PhyML program (Guindon and Gascuel, 2003) was used to estimate the unrooted gene trees. After that is simulated, the ABC-algorithm is applies to the five taxa of DNA sequences to calculate the posterior probability using topologies and splits by using the same method used for five taxa with known gene trees. Figure 2.8 shows the species trees with branch lengths that this study uses.

Table 2.6: Posterior Topology for Five Taxa with DNA sequences

Type of Tree	Species Tree (x, y, z)	Prop. Correct	Coverage Probability	Average of the Posterior Probability							# of trees in 90% CR.
				1	2	3	4	5	6	7	
Caterpillar	(0.1,0.1,1.0)	100	100	60.1	21.5	1.8	1.9	7.2	5	2.5	3.3
Balance	(0.1,0.1,1.0)	0.35	100	6.4	7	21.4	20.4	19.6	20.8	4.4	5.4
Pseudo-caterpillar	(0.1,0.1,1.0)	100	100	8.1	6.3	5.9	7.8	61.3	5.3	5.3	4.9

2.4.1 Result

Tables 2.6 and 2.7 show the posterior probability using topologies and splits for species trees with DNA sequences. From both tables, the present study displays that the coverage probability for containing the correct species in all the iterations for all types of species. Species Tree 2, which is the caterpillar tree with DNA sequences, matches Tree 1, which is the correct tree 100% of the time, in the computation of both the species topology and the species split. But the average posterior probability for Species 2 in both the topology and split with DNA sequences is less than the average probability for both the topology counts and the split counts without DNA sequences, as shown in Table 2.4, Table 2.5, Table 2.6 and Table 2.7. Moreover, there is no significant correlation between the topology and split for Species 2 with DNA sequences where the p-value is 0.687, since the present study does it with a small sample size, which is 10 iterations.

For species tree 8, which is a balanced tree, the proportion of matches for the correct tree, which is Tree 6, is 35% with DNA sequences. For the split, the propor-

Table 2.7: Posterior Split for Five Taxa with DNA sequences

Type of Tree	Species Tree (x, y, z)	Prop. Correct	Coverage Probability	Average of the Posterior Probability							# of trees in 90% CR.
				1	2	3	4	5	6	7	
Caterpillar	(0.1,0.1,1.0)	100	100	49.5	21.2	4.9	4.9	11.3	5.1	2.9	4.4
Balance	(0.1,0.1,1.0)	0.2	100	6.9	7.5	19.7	19.6	22.2	19.8	4.3	5.3
Pseudo-caterpillar	(0.1,0.1,1.0)	100	100	11	9.1	8.6	9.6	51.1	5.1	5.5	5.2

tion of correct trees for Species Tree 8, which is computed with a DNA sequence, is less than the proportion of the correct tree for Species 8 without DNA, as seen in Table 2.5 and Table 2.7. Also, the average posterior probability for Species Tree 8 in both the topology and the split with DNA sequences is less than the average probability for both the topology and the split without DNA sequences, as shown in Tables 2.4, 2.5, 2.6 and 2.7. Moreover, there is no significant correlation between the topology and the split for Species Tree 8 with DNA sequences, where the p-value is 0.3845, since the present study does it with a small sample size, which is 10 iterations.

Species Tree 13 (a pseudocaterpillar shape) with DNA sequences matches the correct tree by 100% of the time; it is also had coverage probability of 100% in both topology counts and split counts. Moreover, the average of posterior probability is the highest in both topology counts and split counts by 61.30 and 51.10 respectively. There is no significant correlation between the topology and the split for Species 13 with DNA sequences, where the p-value is 0.067, since the present study does it with a small sample size, which is 10 iterations. For these examples, inference of the caterpillar species tree and pseudocaterpillar species tree were improved using estimated

rather than known gene trees, while inference of the balanced species tree was somewhat worse for estimated gene trees. Although estimated rather than known gene trees typically make species tree inference more difficult (Huang et al., 2010; Roch and Warnow, 2015), it is possible to speculate that biases in estimated gene trees (Huelsenbeck and Kirkpatrick, 1996) might lead to a gene tree distribution which favors some trees at the expense of others, and to note that improved performance in the anomaly zone using estimated gene trees rather than known gene trees has been observed previously (Wang and Degnan, 2011).

2.5 Equal Branch length for 5-taxa

This section discusses applying the ABC method to a species tree with equal branch lengths in order to investigate inferring the species tree. Also, it investigated the effect of having significantly longer branches, as well as having a star-shaped species tree with all internal branches having length 0. The caterpillar and balanced tree topologies were simulated with $x = y = z = 0$, 0.1, 0.5, 1.0, 2.0, and 3.0 using the same settings as before, with $J = 50,000$, and to retain the best 100 proposed species trees, with 50 replicates for each combination of topology and branch lengths. When all branches have the same length, and are long in coalescent units, there is less variation in the gene trees. In this case, a large number of gene trees have the unrooted topology $((a, b), c, (d, e))$, which matches the unrooted topology of all of the species trees in the prior. If all gene trees have this topology, then there should be little or no information in the data to determine the correct rooted species tree. In this case, it is possible to expect that the highest posterior probability tree would be equally likely to be any of the seven trees in the prior.

Table 2.8: Posterior Topology for Five Taxa Equal Branch Length

Type of Tree	Species Tree (x, y, z)	Prop. Correct	Coverage Probability	Average of the Posterior Probability							# of trees in 90% CR.
				1	2	3	4	5	6	7	
Caterpillar	(0.0,0.0,0.0)	0.27	0.64	21.24	0.00	21.22	24.44	20.14	8.52	4.32	4.32
	(0.1,0.1,0.1)	0.76	0.96	53.62	23.96	2.40	2.60	7.26	7.46	2.70	3.04
	(0.5,0.5,0.5)	100	100	64.70	0.00	0.00	0.02	2.82	31.5	0.96	2.06
	(1.0,1.0,1.0)	0.36	0.94	25.34	25.32	2.10	2.82	9.46	28.08	6.88	4.14
	(2.0,2.0,2.0)	0.00	0.46	10.64	0.00	8.96	8.74	9.76	42.5	19.38	4.92
	(3.0,3.0,3.0)	0.00	0.14	8.08	0.00	7.30	7.64	7.04	45.94	23.98	4.62
Balance	(0.0,0.0,0.0)	0.02	0.34	18.48	0.00	22.62	24.00	21.78	8.70	4.16	4.30
	(0.1,0.1,0.1)	0.02	0.50	30.36	28.18	5.54	4.8	16.46	10.82	3.84	3.76
	(0.5,0.5,0.5)	100	100	30.54	0.00	0.98	1.40	8.10	55.76	3.22	2.68
	(1.0,1.0,1.0)	0.65	0.92	15.94	14.82	8.52	8.08	13.02	28.74	10.88	5.26
	(2.0,2.0,2.0)	100	100	7.86	0.00	8.10	8.32	7.64	53.74	14.34	4.7
	(3.0,3.0,3.0)	100	100	6.92	0.00	6.20	7.06	6.14	53.10	20.58	4.32

2.5.1 Result

According to this result in Table 2.8 and Table 2.9, it is difficult to infer the caterpillar tree with long branch lengths. When the branch length in the balanced tree is long, the highest posterior probability is obtained in each case of the topology and each case of the split, and the highest proportion of correct trees is obtained. Both species with caterpillar shape and balanced shape produce very accurate inferences when the branch length is (0.5, 0.5, 0.5) with 100 loci. Also, when the branch length is between zero and one, the level of accuracy to infer species trees has variation.

Figure 2.9 shows that the accuracy for inferring the caterpillar tree increases

Table 2.9: Posterior Split for Five Taxa Equal Branch Length

Type of Tree	Species Tree (x, y, z)	Prop. Correct	Coverage Probability	Average of the Posterior Probability							# of trees in 90% CR.
				1	2	3	4	5	6	7	
Caterpillar	(0.0,0.0,0.0)	0.17	0.88	20.90	0.00	21.90	25.74	23.24	5.28	2.80	4.20
	(0.1,0.1,0.1)	0.76	0.94	47.96	24.74	4.46	4.04	10.28	6.10	2.42	3.46
	(0.5,0.5,0.5)	0.91	100	59.34	0.00	0.02	0.00	5.02	34.40	1.22	2.18
	(1.0,1.0,1.0)	0.33	0.94	24.46	24.62	2.04	2.54	11.34	26.88	8.12	4.20
	(2.0,2.0,2.0)	0.00	0.54	11.10	0.00	8.88	8.96	9.22	42.24	19.60	4.88
	(3.0,3.0,3.0)	0.00	0.20	8.16	0.00	7.22	7.70	7.10	46.62	23.18	4.54
Balance	(0.0,0.0,0.0)	0.00	0.08	20.40	0.00	23.34	24.28	23.86	5.36	2.54	4.08
	(0.1,0.1,0.1)	0.00	0.34	27.78	26.6	7.06	7.6	19	8.86	3.1	4.14
	(0.5,0.5,0.5)	0.99	100	29.82	0.00	1.34	1.64	9.50	53.70	3.98	2.94
	(1.0,1.0,1.0)	0.72	0.92	15.76	15.26	9.08	8.08	12.06	28.84	10.92	5.4
	(2.0,2.0,2.0)	100	100	7.26	0.00	7.82	7.92	7.48	55.50	14.02	4.54
	(3.0,3.0,3.0)	100	100	6.96	0.00	5.94	6.76	6.20	54.18	19.96	4.30

when the branch length increases until the branch reaches the length (0.5, 0.5, 0.5), which gives the highest proportion of correct trees to match the true tree and then starts decreasing until the branch reaches the length (2, 2, 2). Then, it goes to zero, which means that it is difficult to infer the caterpillar shape of a species tree with long branch lengths. Figure 2.9 shows the species tree with a balanced shape, which matches the long branch length. However, when the branch length is between zero and one the proportion of correct trees goes to the highest branch length (0.5, 0.5, 0.5).

For the caterpillar topology, it is observed that, for long branches, the proportion of times that the correct species tree is inferred is highest when $x = y = z = 0.5$,

and quickly goes to 0 when the branch length is 2 or more coalescent units. When the branch lengths are 0 (i.e. a star tree), the proportion of times that the caterpillar is inferred is higher than the chance value of 1/7, with a proportion of roughly 26% using topology counts (Fig. 2.9). This suggests that the prior is informative. Similarly, having 0% chance of recovering the correct tree for long branches suggests an informative prior. For these cases, the prior, although uniform for the topology, also includes a prior for branch lengths which is not well suited to the data.

Caterpillar species trees tend to have higher gene-tree discordance than balanced species trees, given similar branch lengths (Degnan and Salter, 2005), which could explain why the caterpillar is favored in the posterior when gene trees are purely random (the star species tree) and why caterpillars are under-estimated when the gene trees have no variation (long species tree branches). Consistent with this prediction, the proportion of times the balanced species tree is inferred is lower than would be expected for an uninformative prior for the star tree, and there is a bias in favor of balanced topologies when there is no variation in the gene tree topologies (Fig. 2.9). These examples also illustrate that, unlike the case of inferring rooted species trees from rooted gene trees, or unrooted species trees from unrooted gene trees, longer internal branches do not necessarily make the inferences easier. Inference of the rooted species tree from unrooted gene trees requires variation in the gene trees.

2.6 Eight Taxa

This section discusses how more than 5 taxa work and can be generalized for more taxa. This section contains two subsections, which are the simulation of 8 taxa and the result.

2.6.1 Simulation of 8-Taxon trees

A slightly modified ABC method (Algorithm 3) is used to compute the posterior probability when eight-taxon trees were inferred. The ape library (Paradis et al., 2004) and the sets library (Hornik and Meyer, 2009) from the R-package (Ihaka and Gentleman, 1996) are used to do the calculation of the distance between the simulated data and the observed data, which is done as the first step to compute the difference between the observed data and the simulated data and then the inverse process by calculating the difference between the simulated data and the observed data. This process works to find the elements that are in the observed data but not in the simulated data and inversely to find the elements in the simulated data but not in the observed data. After all this process, it is necessary to find the union of the differences; then it is necessary to sum the number of elements in the union set. All these steps are done depending on the splits of the tree. Two types of 8-taxon trees were used, which are the caterpillar tree and the balanced tree. Figure 2.10 below shows that the first two shapes with 8 taxa in the left are a rooted and an unrooted caterpillar tree, and the second two shapes of 8 taxa in the right are a rooted and an unrooted balanced tree.

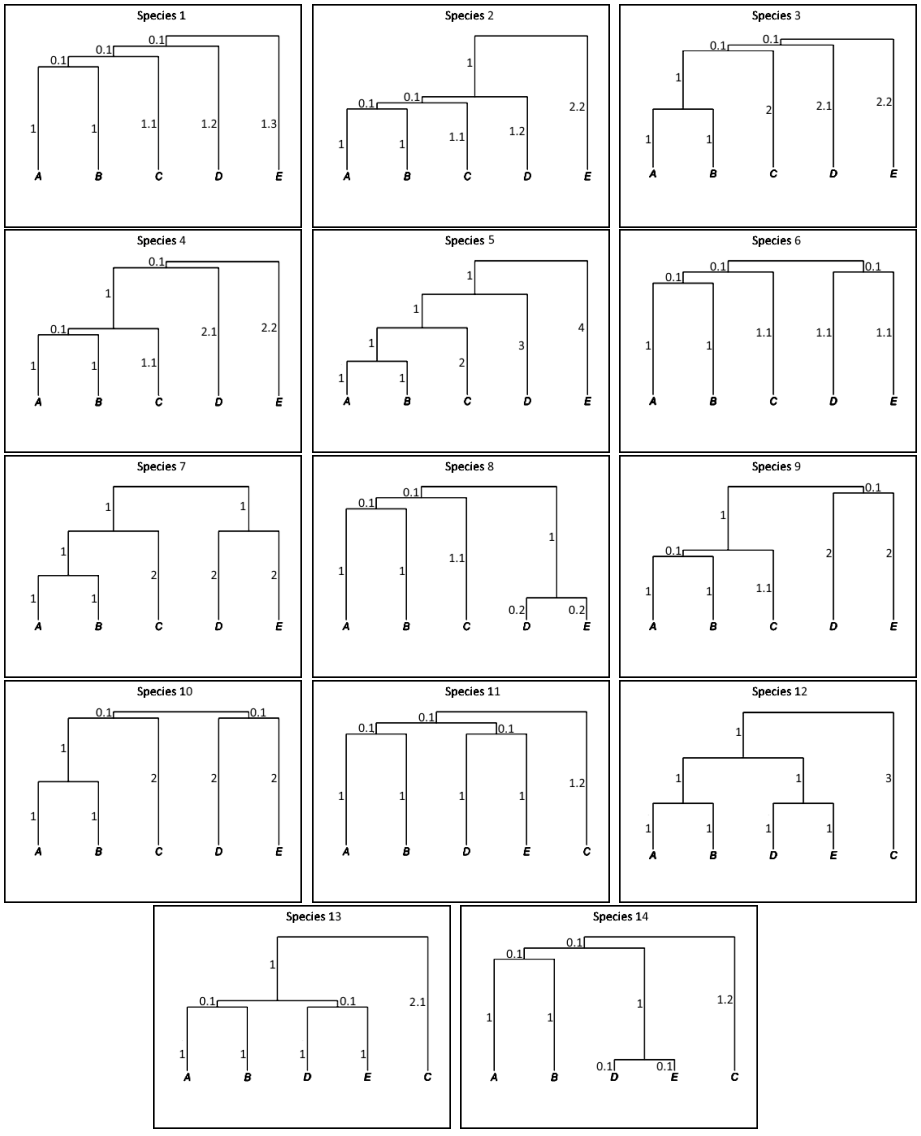

Figure 2.3: Species Tree Shapes and Branch Length

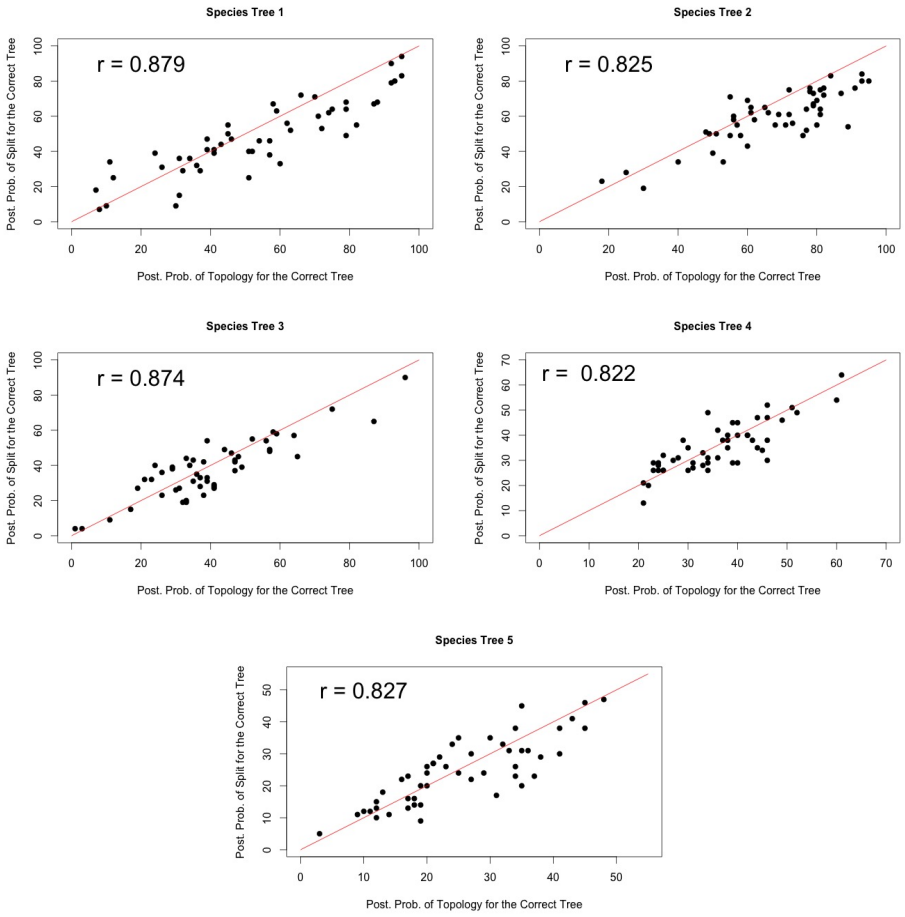

Figure 2.4: Correlation for Caterpillar Trees

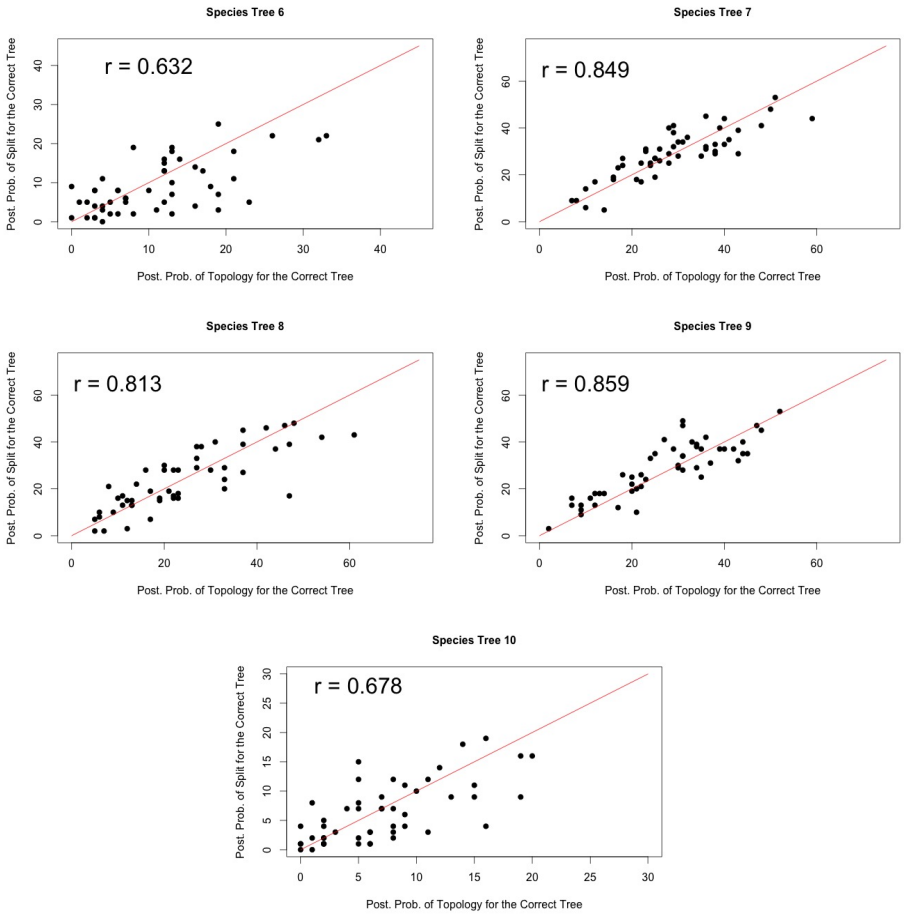

Figure 2.5: Correlation for Balanced Trees

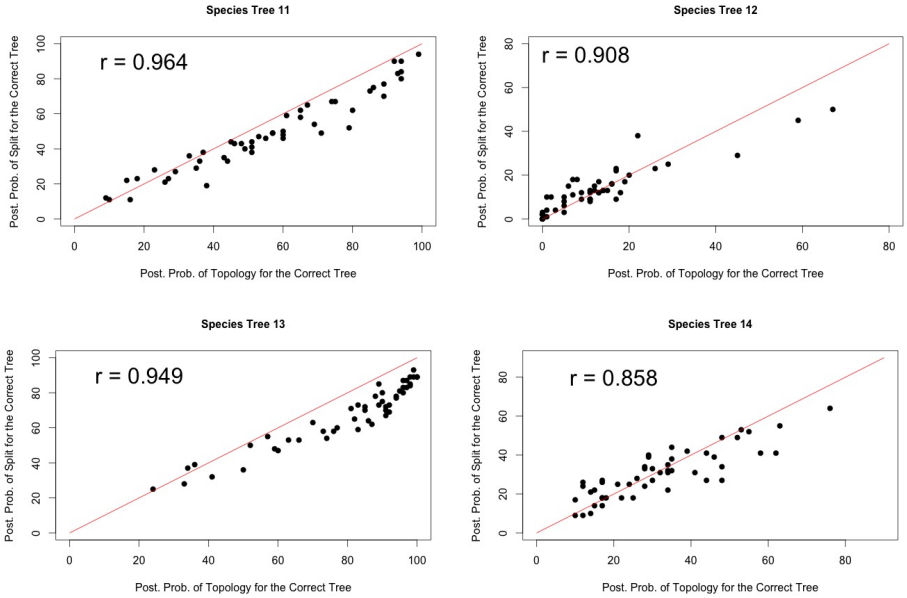

Figure 2.6: Correlation for Pseudocaterpillar Trees

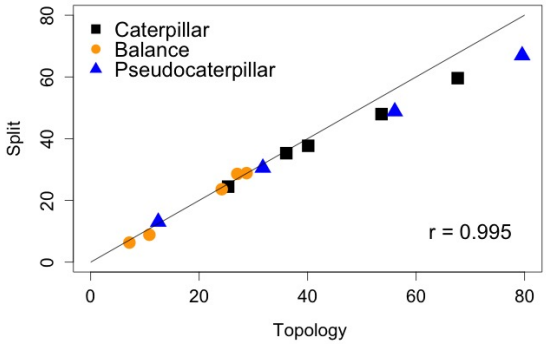

Figure 2.7: Correlation of Average Posterior Probability for topology and Split

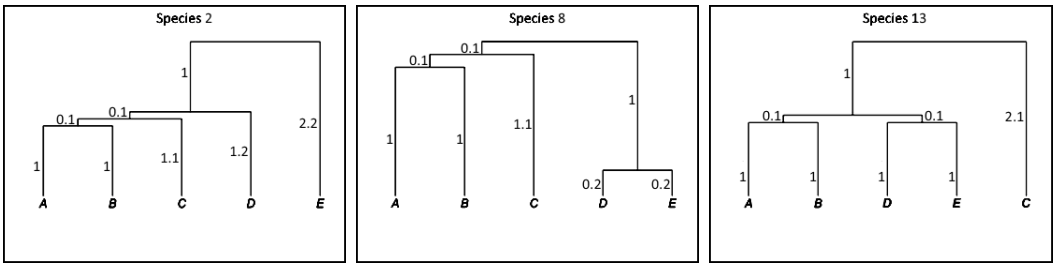

Figure 2.8: Species Tree Shapes and Branch Length to Simulate DNA Sequenaces

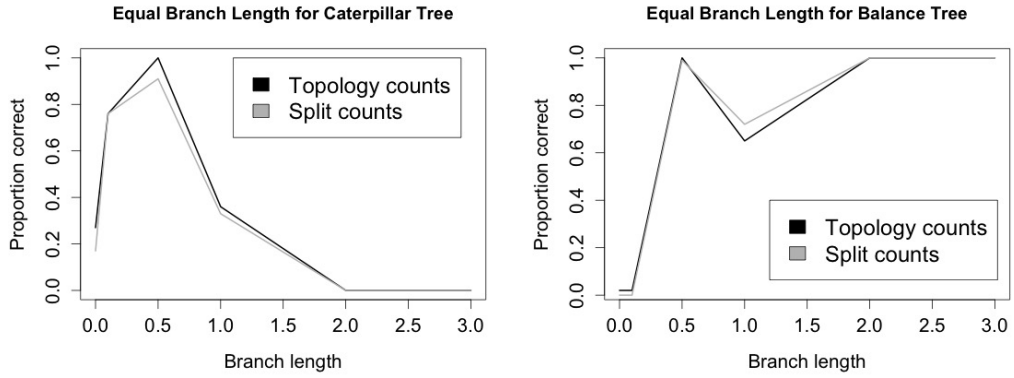

Figure 2.9: Caterpillar and Balance Trees with Equal Branch Length

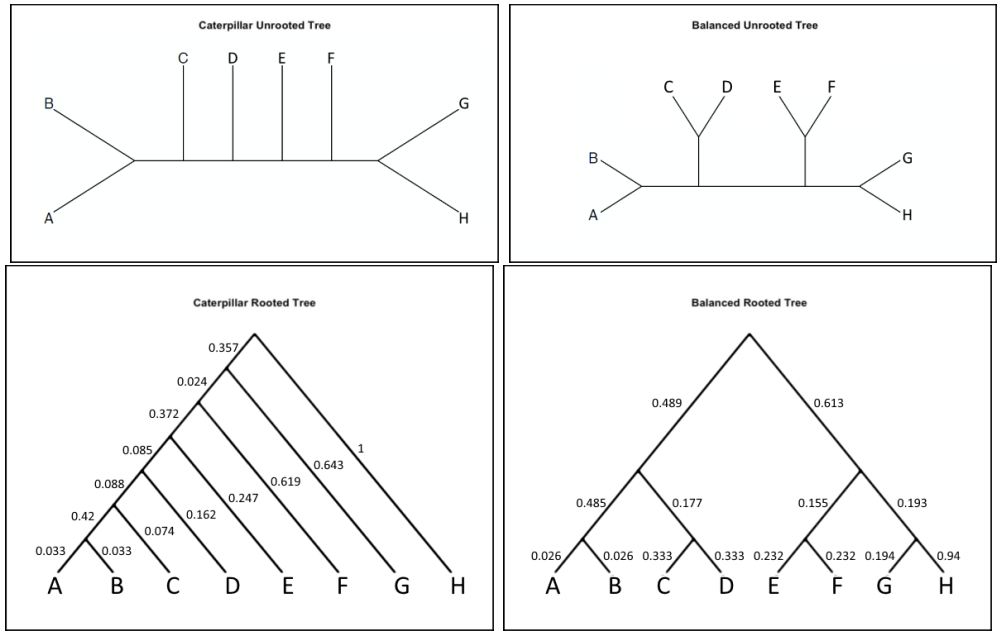

Figure 2.10: Rooted tree vs. unrooted tree for 8 taxa

2.6.2 Result

Table 2.10: Average Posterior probability for 8-Taxa With Caterpillar Tree.

posterior of Topology	RF dist	AB	ABC	ABCD	ABC DE	ABC DEF	ABC DEFG	GH	Coverage of Probability	# of Trees
24.46	2	98.22	96.38	93.08	82.48	54.74	24.48	51.66	100	6.16

To calculate the posterior probability, the split program is used. The posterior probability of the topology is zero matching the caterpillar tree for 8 taxa. Since the tree distance was constant; in this case, it was always equal to two. As seen in Table 2.10, since the highest posterior probability tree never matched the species tree, but instead always had a (G,H) clade with all other clades being correct. As seen in Table 2.12, the posterior probability in balanced trees is much flatter than it is for the caterpillar tree. The reason behind this is that the number of trees, which equals 90%, is usually 10 out of 13 trees of the prior, but in the caterpillar tree, it is usually 6 out of 13 trees of the prior. A balanced tree is always obtained in all attempts. Also, there is no difference between the original tree and the tree generated by the program.

Table 2.12: Average Posterior probability for 8-Taxa With Balance Tree.

posterior of Topology	RF dist	AB	CD	ABCD	EF	GH	EFGH	Coverage of Probability	# of Trees
24.68	0.00	83.48	81.98	55.60	86.78	87.18	69.08	100	9.58

Figure 2.11 shows that the average posterior probability of each clade of the caterpillar tree that is acquired from the consensus program. It also shows the shape

that is acquired from the consensus program and also shows how it is different from the caterpillar tree. Moreover, it also shows the average posterior probability of the tree for each clade, which was acquired from the consensus program. Figure 2.12 displays the average posterior probability of each clade of the balanced tree.

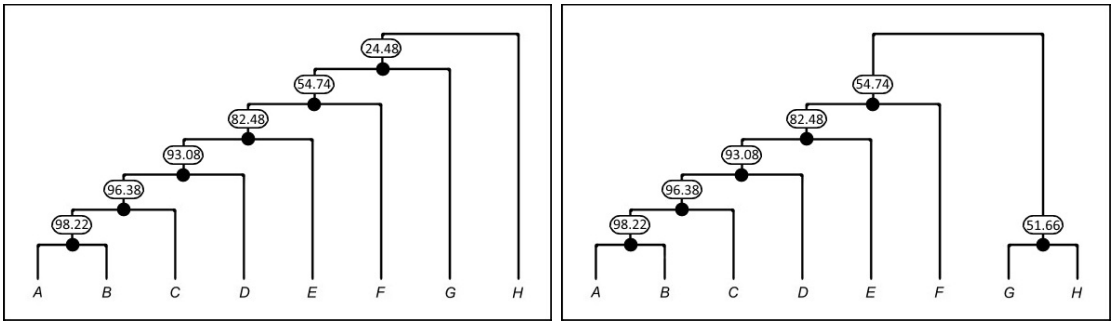

Figure 2.11: Average of The Posterior Probability for Caterpillar Tree and Tree From Consensus Program

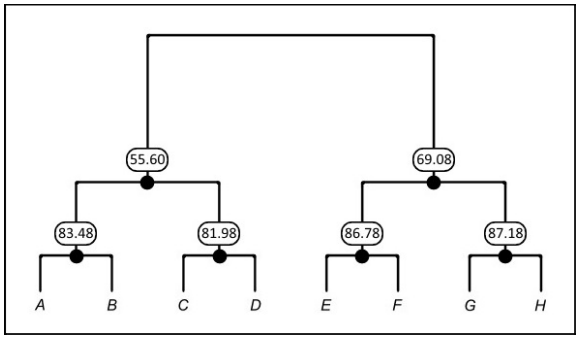

Figure 2.12: Average of The Posterior Probability for Balanced Tree

Chapter 3

Maximum Likelihood Estimate (MLE)

This chapter covers the idea of the maximum likelihood estimate (MLE) of the phylogenetic tree. From among the many methods of the estimation in the statistics field, the MLE is one of the most popular. In order to find the optimal value of the parameters, it is necessary to find the ML, which leads to computing the likelihood function. To compute the MLE, this study, from the outset, needs to compute the likelihood function, which is as follows for unrooted trees, and this is also the formula that the PhyloNet program (Than et al., 2008) uses to compute the MLE for unrooted trees:

$$\prod_{i=1}^{(2m-5)!!} P_i^{n_i} = \prod_{i=1}^{(2m-5)!!} (\sum_{j=1}^{2m-3} P_{ij})^{n_i}$$

Where i is the index to the topology, P_i= Probability of the i^{th} unrooted topology, n_i is the number of trees observed with topology i, j indexes the root location within the i^{th} unrooted topology, and m is the number of taxa. From the left hand side of the equation, the liklihood is multinomial, where the number of categories is the

number of unrooted tree topologies. However, We also use the PhyloNet program to compute the likelihood function for rooted trees, and also the PhyloNet program needs to use this formula to compute the MLE:

$$\prod_{ij} P_{ij}^{n_{ij}} = \prod_{k=1}^{(2m-3)!!} P_k^{n_k}$$

In the following sections, the method of the MLE is used, as are the results obtained from it.

3.1 MLE Method

Allman et al. (2011b) provided the three distributions for three particular trees based on the three tree shapes for 5-taxon trees (caterpillar, pseudocaterpillar, and balanced) of unrooted gene trees. The present work gathers all those three distributions of unrooted gene trees here. The remaining distributions of unrooted trees needed for all seven possible rootings of the unrooted tree $((a, b), c, (d, e))$ are then computed. The first type for the distribution of unrooted gene trees is taken from a 5 taxa Pseudocaterpillar and there is only one Pseudocaterpillar needed. The rooted caterpillar species tree is considered the second type for the distribution of unrooted gene trees and it is left with three rooted caterpillar species trees, which are calculated. In addition, the third type of distribution of unrooted gene trees is balanced species tree, which has two rooted balanced species trees, three of them collected from Allman et al. (2011b) and another four based on them, which are new to the present work.

3.1.1 Rooted Caterpillar Species Tree

The following 5-taxon rooted caterpillar species tree is the first equation in Allman et al. (2011b):

σ^+ = **((((a,b):x,c):y,d):z,e)**,

Let $X = exp(-x)$, $Y = exp(-y)$, and $Z = exp(-z)$. Then the distribution of unrooted gene tree T_i under the coalescent is given by u_i=$P_{\sigma^+}(T_i)$ with **The first species**:

$u_1 = 1 - \frac{2}{3}\mathrm{X} - \frac{2}{3}\mathrm{Y} + \frac{1}{3}\mathrm{XY} + \frac{1}{18}\mathrm{XY}^3 + \frac{1}{90}\mathrm{XY}^3Z^6$,

$u_2 = \frac{1}{3}\mathrm{Y} - \frac{1}{6}\mathrm{XY} - \frac{1}{9}\mathrm{XY}^3 + \frac{1}{90}\mathrm{XY}^3Z^6$,

$u_3 = \frac{1}{3}\mathrm{Y} - \frac{1}{6}\mathrm{XY} - \frac{1}{18}\mathrm{XY}^3 - \frac{2}{45}\mathrm{XY}^3Z^6$,

$u_4 = u_{13} = \frac{1}{3}\mathrm{X} - \frac{1}{3}\mathrm{XY} + \frac{1}{18}\mathrm{XY}^3 + \frac{1}{90}\mathrm{XY}^3Z^6$,

$u_5 = u_{12} = \frac{1}{6}\mathrm{XY} - \frac{1}{9}\mathrm{XY}^3 + \frac{1}{90}\mathrm{XY}^3Z^6$,

$u_6 = u_9 = \frac{1}{6}\mathrm{XY} - \frac{1}{18}\mathrm{XY}^3 - \frac{2}{45}\mathrm{XY}^3Z^6$,

$u_7 = u_8 = u_{10} = u_{11} = u_{14} = u_{15} = \frac{1}{18}\mathrm{XY}^3 + \frac{1}{90}\mathrm{XY}^3Z^6$.

The following 5-taxon-rooted caterpillar species tree is new offered in the present work:

σ^+ = **((((a,b):x,c):y,e):z,d)**,

Under the same assumption of X, Y, and Z, its distribution should be as follows:

$u_1 = 1 - \frac{2}{3}\mathrm{X} - \frac{2}{3}\mathrm{Y} + \frac{1}{3}\mathrm{XY} + \frac{1}{18}\mathrm{XY}^3 + \frac{1}{90}\mathrm{XY}^3Z^6$,

$u_2 = \frac{1}{3}\mathrm{Y} - \frac{1}{6}\mathrm{XY} - \frac{1}{18}\mathrm{XY}^3 - \frac{2}{45}\mathrm{XY}^3Z^6$,

$u_3 = \frac{1}{3}\mathrm{Y} - \frac{1}{6}\mathrm{XY} - \frac{1}{9}\mathrm{XY}^3 + \frac{1}{90}\mathrm{XY}^3Z^6$,

$u_4 = u_{13} = \frac{1}{3}\mathrm{X} - \frac{1}{3}\mathrm{XY} + \frac{1}{18}\mathrm{XY}^3 + \frac{1}{90}\mathrm{XY}^3Z^6$,

$u_5 = u_{12} = \frac{1}{6}\mathrm{XY} - \frac{1}{18}\mathrm{XY}^3 - \frac{2}{45}\mathrm{XY}^3Z^6$,

$u_6 = u_9 = \frac{1}{6}\mathrm{XY} - \frac{1}{9}\mathrm{XY}^3 + \frac{1}{90}\mathrm{XY}^3Z^6$,

$u_7 = u_8 = u_{10} = u_{11} = u_{14} = u_{15} = \frac{1}{18}\mathrm{XY}^3 + \frac{1}{90}\mathrm{XY}^3Z^6$.

The following 5-taxon-rooted caterpillar species tree is also new offered in the present work:

σ^+ = **((((d,e):x,c):y,a):z,b)**,

Under the same assumption of X, Y, and Z, its distribution should be as follows:

$u_1 = 1 - \frac{2}{3}X - \frac{2}{3}Y + \frac{1}{3}XY + \frac{1}{18}XY^3 + \frac{1}{90}XY^3Z^6$,

$u_2 = u_3 = \frac{1}{3}X - \frac{1}{3}XY + \frac{1}{18}XY^3 + \frac{1}{90}XY^3Z^6$,

$u_4 = \frac{1}{3}Y - \frac{1}{6}XY - \frac{1}{18}XY^3 - \frac{2}{45}XY^3Z^6$,

$u_5 = u_6 = u_8 = u_9 = u_{11} = u_{12} = \frac{1}{18}XY^3 + \frac{1}{90}XY^3Z^6$,

$u_7 = u_{10} = \frac{1}{6}XY - \frac{1}{18}XY^3 - \frac{2}{45}XY^3Z^6$,

$u_{13} = \frac{1}{3}Y - \frac{1}{6}XY - \frac{1}{9}XY^3 + \frac{1}{90}XY^3Z^6$,

$u_{14} = u_{15} = \frac{1}{6}XY - \frac{1}{9}XY^3 + \frac{1}{90}XY^3Z^6$.

The following 5-taxon-rooted caterpillar species tree is the final new distribution derived for the caterpillar shape offered in the present work:

σ^+ = **((((d,e):x,c):y,b):z,a)**,

Under the same assumption of X, Y, and Z, its distribution should be as follows:

$u_1 = 1 - \frac{2}{3}X - \frac{2}{3}Y + \frac{1}{3}XY + \frac{1}{18}XY^3 + \frac{1}{90}XY^3Z^6$,

$u_2 = u_3 = \frac{1}{3}X - \frac{1}{3}XY + \frac{1}{18}XY^3 + \frac{1}{90}XY^3Z^6$,

$u_4 = \frac{1}{3}Y - \frac{1}{6}XY - \frac{1}{9}XY^3 + \frac{1}{90}XY^3Z^6$,

$u_5 = u_6 = u_8 = u_9 = u_{11} = u_{12} = \frac{1}{18}XY^3 + \frac{1}{90}XY^3Z^6$,

$u_7 = u_{10} = \frac{1}{6}XY - \frac{1}{9}XY^3 + \frac{1}{90}XY^3Z^6$,

$u_{13} = \frac{1}{3}Y - \frac{1}{6}XY - \frac{1}{18}XY^3 - \frac{2}{45}XY^3Z^6$,

$u_{14} = u_{15} = \frac{1}{6}XY - \frac{1}{18}XY^3 - \frac{2}{45}XY^3Z^6$.

3.1.2 Pseudocaterpillar Species Tree

The following 5-taxon-rooted pseudocaterpillar species tree is the second equation in Allman et al. (2011b):

σ^+ = **(((a,b):x,(d,e):y):z,c)**,

Under the same assumption of X, Y, and Z, its distribution should be as follows:

$u_1 = 1$ - $\frac{2}{3}$X - $\frac{2}{3}$Y + $\frac{4}{9}$XY - $\frac{2}{45}$XYZ^6,
$u_2 = u_3 = \frac{1}{3}$Y - $\frac{5}{18}$XY + $\frac{1}{90}$XYZ^6,
$u_4 = u_{13} = \frac{1}{3}$X - $\frac{5}{18}$XY + $\frac{1}{90}$XYZ^6,
$u_5 = u_6 = u_7 = u_9 = u_{10} = u_{12} = u_{14} = u_{15} = \frac{1}{18}$XY + $\frac{1}{90}$XYZ^6,
$u_8 = u_{11} = \frac{1}{9}$XY - $\frac{2}{45}$XYZ^6.

3.1.3 Balanced Species Tree

The following 5-taxon-rooted balance species tree is the third equation in Allman et al. (2011b):
σ^+ = **(((a,b):x,c):y,(d,e):z)**,
Under the same assumption of X, Y, and Z, its distribution should be as follows:
$u_1 = 1$ - $\frac{2}{3}$X - $\frac{2}{3}$YZ + $\frac{1}{3}$XYZ + $\frac{1}{15}$XY^3Z,
$u_2 = u_3 = \frac{1}{3}$YZ - $\frac{1}{6}$XYZ - $\frac{1}{10}$XY^3Z,
$u_4 = u_{13} = \frac{1}{3}$X - $\frac{1}{3}$XYZ + $\frac{1}{15}$XY^3Z,
$u_5 = u_6 = u_9 = u_{12} = \frac{1}{6}$XYZ - $\frac{1}{10}$XY^3Z,
$u_7 = u_8 = u_{10} = u_{11} = u_{14} = u_{15} = \frac{1}{15}XY^3$Z.

The following 5-taxon-rooted balance species tree is new offered in the present work:
σ^+ = **(((d,e):x,c):y,(a,b):z)**,
Under the same assumption of X, Y, and Z, its distribution should be as follows:
$u_1 = 1$ - $\frac{2}{3}$X - $\frac{2}{3}$YZ + $\frac{1}{3}$XYZ + $\frac{1}{15}$XY^3Z,
$u_2 = u_3 = \frac{1}{3}$X - $\frac{1}{3}$XYZ + $\frac{1}{15}$XY^3Z,
$u_4 = u_{13} = \frac{1}{3}$YZ - $\frac{1}{6}$XYZ - $\frac{1}{10}$XY^3Z,
$u_5 = u_6 = u_8 = u_9 = u_{11} = u_{12} = \frac{1}{15}XY^3$Z.
$u_7 = u_{10} = u_{14} = u_{15} = \frac{1}{6}$XYZ - $\frac{1}{10}$XY^3Z.

3.2 MLE and Bootstrapping Simulation for Five Taxa without DNA Sequences

The present work uses the R-package to create fifteen functions to calculate the probability for each of the fifteen unrooted gene tree topologies. Each function contains seven trees dependent on five species that have seven shapes of rooted genes. After that, it is necessary to create the log likelihood function, which is the sum of all fifteen functions with respect to the tree, and a crucial step involves using "n.topo.obs" as real data. We use a grid search for $X, Y, Z \in [0, 1]$, corresponding to $x, y, z \in [0, \infty]$ to maximize the function to find the best maximum likelihood estimate for the tree with the best branch length. Moreover, the present work uses the R-package to generate 50 iterations of bootstrapping. To do the bootstrap, the unrooted gene tree topologies are bootstrapped, generating new frequency counts for topologies and splits. The bootstrap allows an estimate of uncertainty in the maximum likelihood estimate.

3.2.1 Results

Table 3.1: Average of Bootstrapping for Five Taxa

Type of Tree	Species Tree (x, y, z)	Prop. Correct	Average of Bootstrapping						
			1	2	3	4	5	6	7
Caterpillar	(0.1,0.1,0.1)	0.62	26.5	8.22	1.22	0.62	2.3	8.38	2.76
	(0.1,0.1,1.0)	0.82	36.52	6.66	0.42	0.08	1.22	4.22	0.88
	(1.0,0.1,0.1)	0.56	23.46	12.32	0.00	0.00	6.44	2.72	5.06
	(0.1,1.0,0.1)	0.42	18.28	16.24	0.00	0.00	0.00	15.48	0.00
	(1.0,1.0,1.0)	0.54	22.14	15.62	0.04	0.08	0.24	11.26	0.62
Balance	(0.1,0.1,0.1)	0.68	9.00	8.5	1.12	1.90	4.82	21.42	3.24
	(1.0,1.0,1.0)	0.24	19.16	17.76	0.32	0.04	0.62	11.64	0.46
	(0.1,0.1,1.0)	0.78	0.04	0.04	3.58	2.30	8.7	32.56	2.78
	(0.1,1.0,0.1)	0.30	17.1	17.58	0.00	0.00	0.02	15.3	0.00
	(1.0,0.1,0.1)	0.40	10.08	11.96	0.00	0.00	7.26	12.78	7.92
Pseudocaterpillar	(0.1,0.1,0.1)	0.80	2.24	2.92	2.36	2.16	29.44	5.70	4.18
	(1.0,1.0,1.0)	0.14	7.88	6.84	6.20	5.24	5.94	11.14	6.76
	(0.1,0.1,1.0)	0.94	1.56	1.68	1.32	1.14	40.94	1.90	1.46
	(0.1,1.0,0.1)	0.42	0.00	0.00	5.82	7.56	19.20	12.32	5.10

Table 3.1 shows the proportion of times when the MLE matched the true tree and the average number of bootstrap replicates that supported a particular tree topology. The first five species trees have a caterpillar tree shape, as seen in Table 3.1. Species Tree 2 has the highest percentage of matching the correct tree according to the MLE, which is 82%. Also, Species Tree 2 has the highest average bootstrap support for the correct species tree, at 36.52 for Tree 1, which is the only correct tree. In contrast, Species Tree 4 has the lowest percentage of matching the correct tree according to the MLE, which is 42%. Also, Species Tree 4 has the lowest average of bootstrapping which supports the MLE, at 18.28 for Tree 1. Thus, the present study shows that both the MLE and the bootstrapping match the correct tree with highest percentage and highest support from the bootstrapping. Moreover, this result displays that the caterpillar Species Tree 2 matches the true tree more than other caterpillar trees.

The second set of five species trees have a balanced tree shape, also in Table 3.1. Species Tree 8 has the highest percentage of matching the correct tree according to the MLE, which is 78%. Also, Species Tree 8 has the highest average bootstrap support for the correct tree, at 32.56 for Tree 6, which is the only correct tree. In contrast, Species Tree 7 has the lowest percentage of matching the correct tree according to the MLE, which is 24%. Also, Species Tree 7 has the lowest average bootstrap support for the correct tree, at 11.64 for Tree 6. Thus, the present study shows that both the MLE and the bootstrapping match the correct tree with highest percentage and highest support from the bootstrapping. Moreover, this result shows that the balanced Species Tree 8 matches the true tree more than other balanced trees.

The last four species trees in Table 3.1 have the pseudocaterpillar shape. Also, the average of bootstrapping supports the MLE tree by 40.94%, which is Species Tree 13. Species Tree 13 obtains the highest proportion of matching the correct MLE tree

(which is Tree 5) by 94%. In contrast, Species Tree 12 obtains the lowest proportion of matching the correct MLE tree by 14%. Species Tree 12 also obtains the lowest average of bootstrapping to support the correct MLE tree by 5.94%. From the result in Table 3.1, the present study finds that all proportions of matching the correct MLE tree and averaging the bootstrapping to support MLE trees are ordered from the highest to the lowest in order to give the same branch length for the caterpillar and pseudocaterpillar shapes. However, the balanced shape has different branch length.

Figure 3.1 shows the correlation between the average bootstrap support, the average posterior probability for topology, and the average posterior probability for splits. These are not strong correlations since they have only a moderate relationship between bootstrapping and the posterior probability. Figure 3.2 and Figure 3.3 show that all caterpillar trees and all balanced trees have a very weak relationship between bootstrapping supporting the MLE tree and the posterior probability matching the correct tree, since each correlation value is very small. However, Species Tree 5 has the highest correlation value between all species of caterpillar shapes and all species of balance shapes. Figure 3.4 shows the pseudocaterpillar trees with a moderate relationship between bootstrapping supporting the MLE tree, the posterior probability for topology counts matching the correct tree, and the posterior probability for split counts matching the correct tree in both Species Tree 12 and Species Tree 14. In Species Tree 13 there is a stronger relationship between the bootstrap support for the MLE tree and the posterior probability for topology counts matching the correct tree, but there is a moderate relationship between the bootstrap support for the MLE tree and the posterior probability for split counts matching the correct tree. Specie Tree 11 has the weakest relationship between all species trees of the pseudocaterpillar shapes. Figure 3.5, Figure 3.6, and Figure 3.7 show the summary of the five numbers for the branch length that is acquired from the simulation for matching a correct tree.

The correlation test for this data is as follows: There is a significant correlation between the bootstrapping supporting the MLE tree and the posterior probability of the topology matching the correct tree. There is also a significant correlation between bootstrap support for the MLE tree and the posterior probability of a split matching the correct tree in Species Tree 5 with the caterpillar shape and Species Trees 12, 13, and 14 with a pseudocaterpillar shape. Moreover, Species Tree 6 with a balanced shape has a significant correlation between the bootstrap support for the MLE tree and the posterior probability of the topology matching the correct tree. However, there is no significant correlation between the bootstrap support for the MLE tree and the posterior probability of a split matching the correct tree. Moreover, there is no significant correlation between the bootstrap support for the MLE tree and the posterior probability of topology matching the correct tree. There is also no significant correlation between the bootstrap support for the MLE tree and the posterior probability of a split matching the correct tree in all of the rest of the species trees.

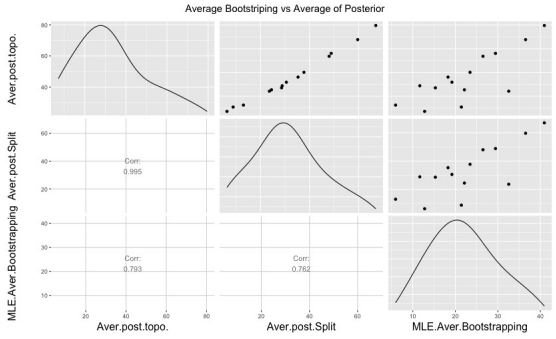

Figure 3.1: Correlation of Average Bootstrapping vs Average Posterior Probability

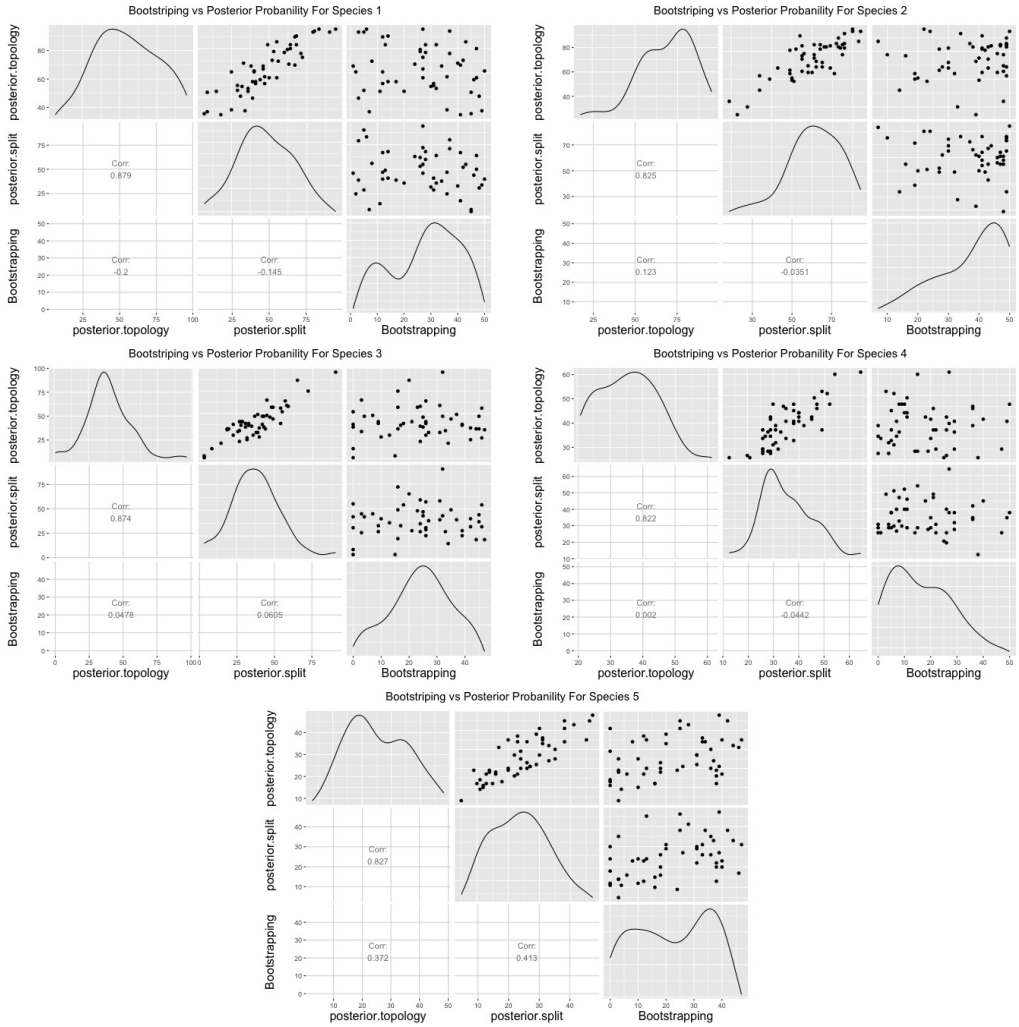

Figure 3.2: Correlation for Caterpillar Trees Bootstrapping vs Posterior Probability

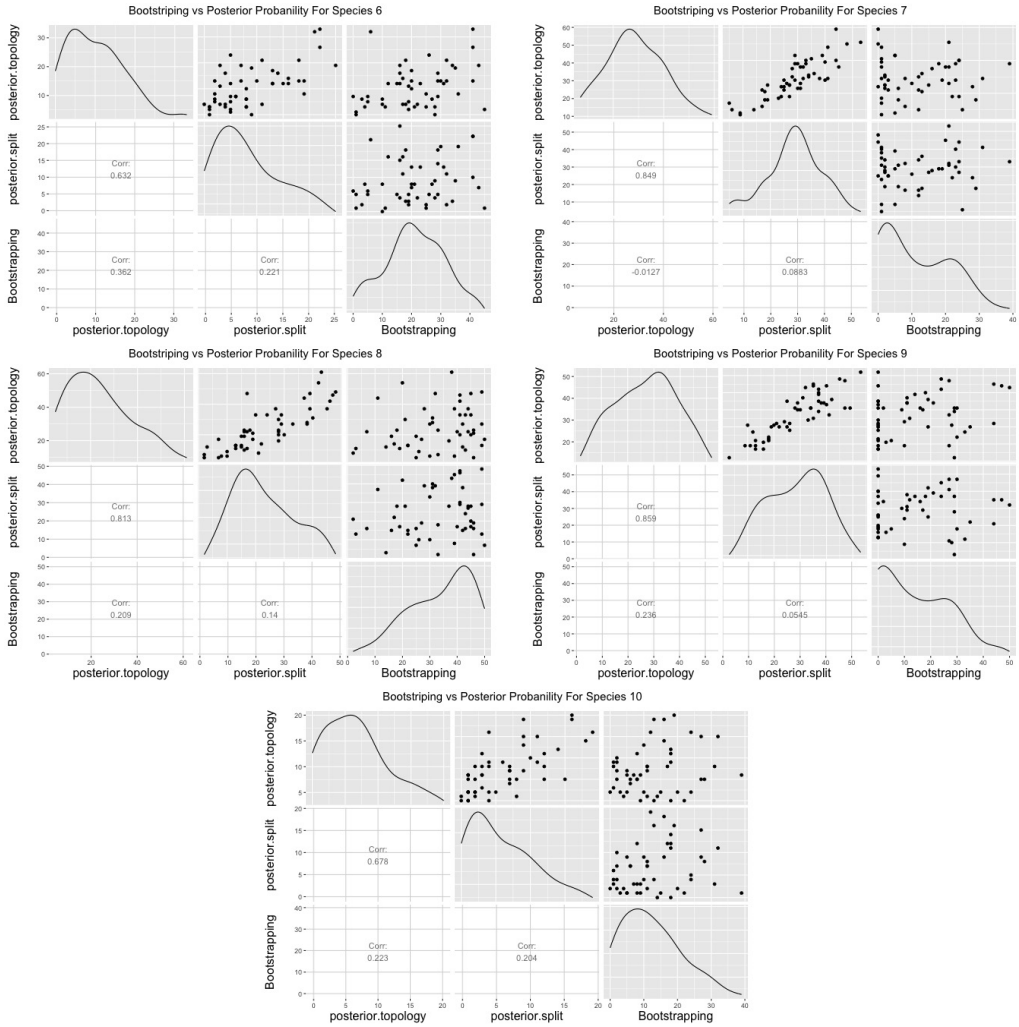

Figure 3.3: Correlation for Balanced Trees Bootstrapping vs Posterior Probability

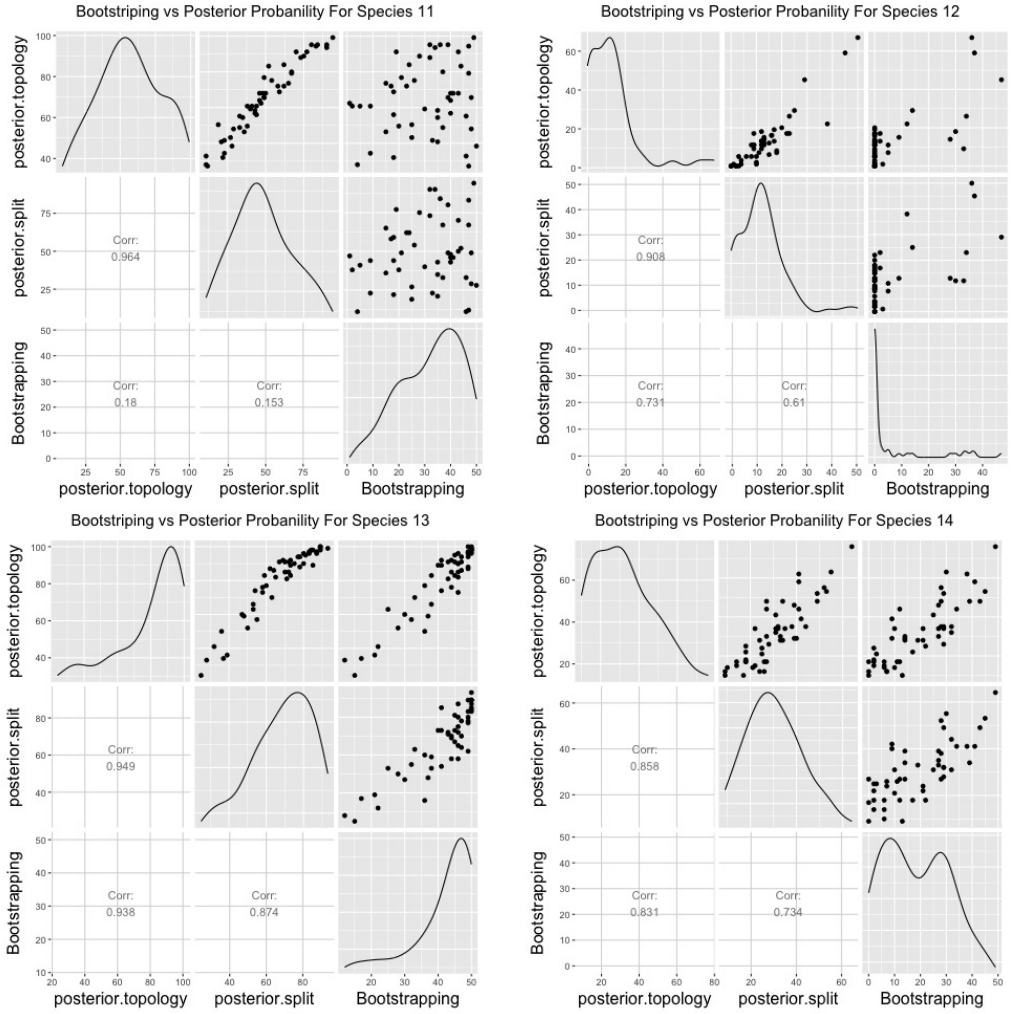

Figure 3.4: Correlation for Pseudocaterpillar Trees Bootstrapping vs Posterior Probability

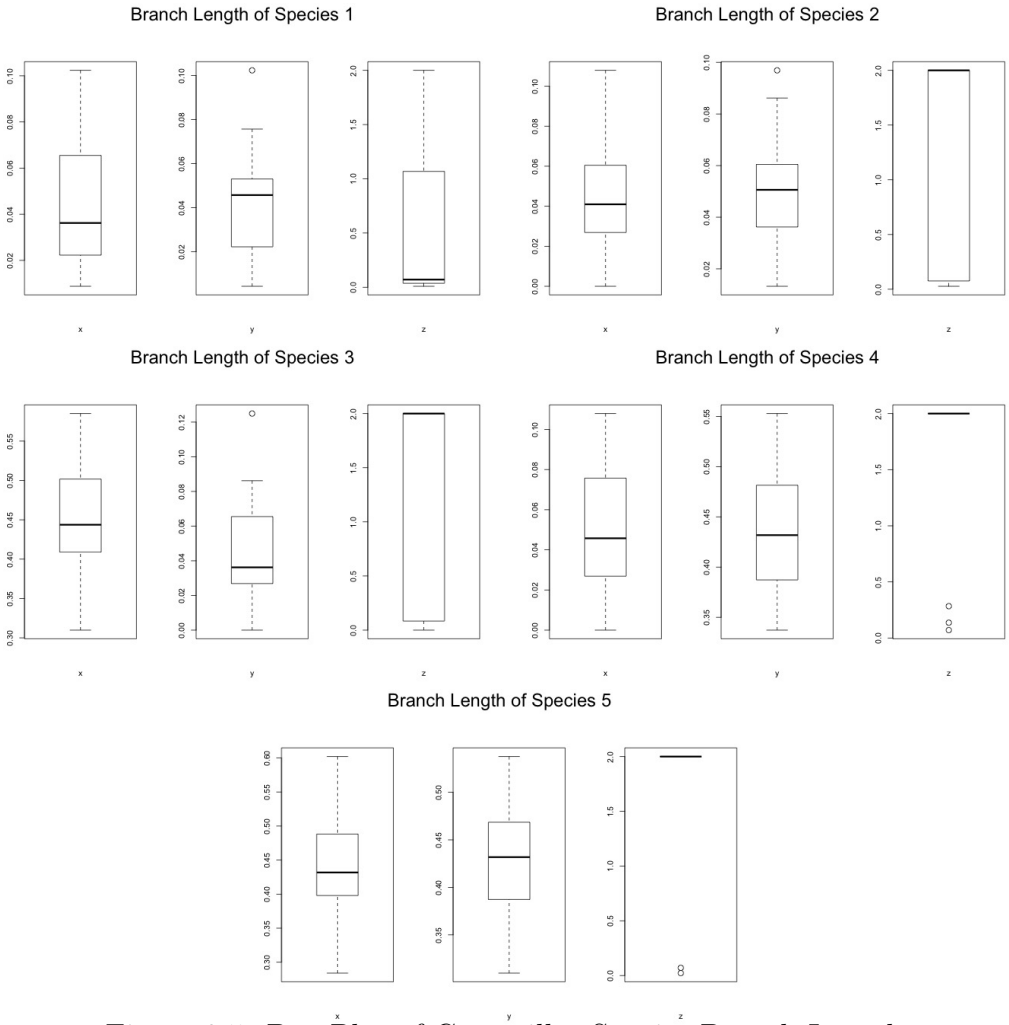

Figure 3.5: Box Plot of Caterpillar Species Branch Length

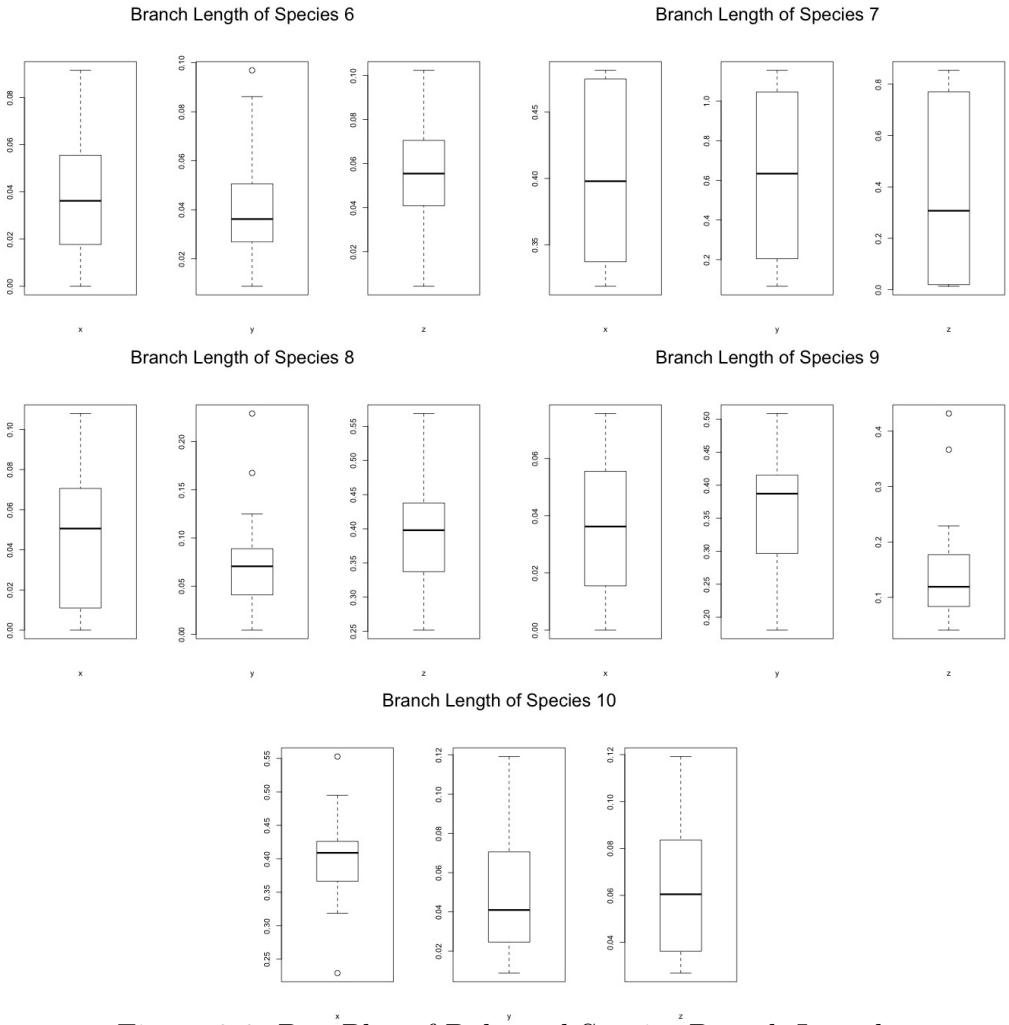

Figure 3.6: Box Plot of Balanced Species Branch Length

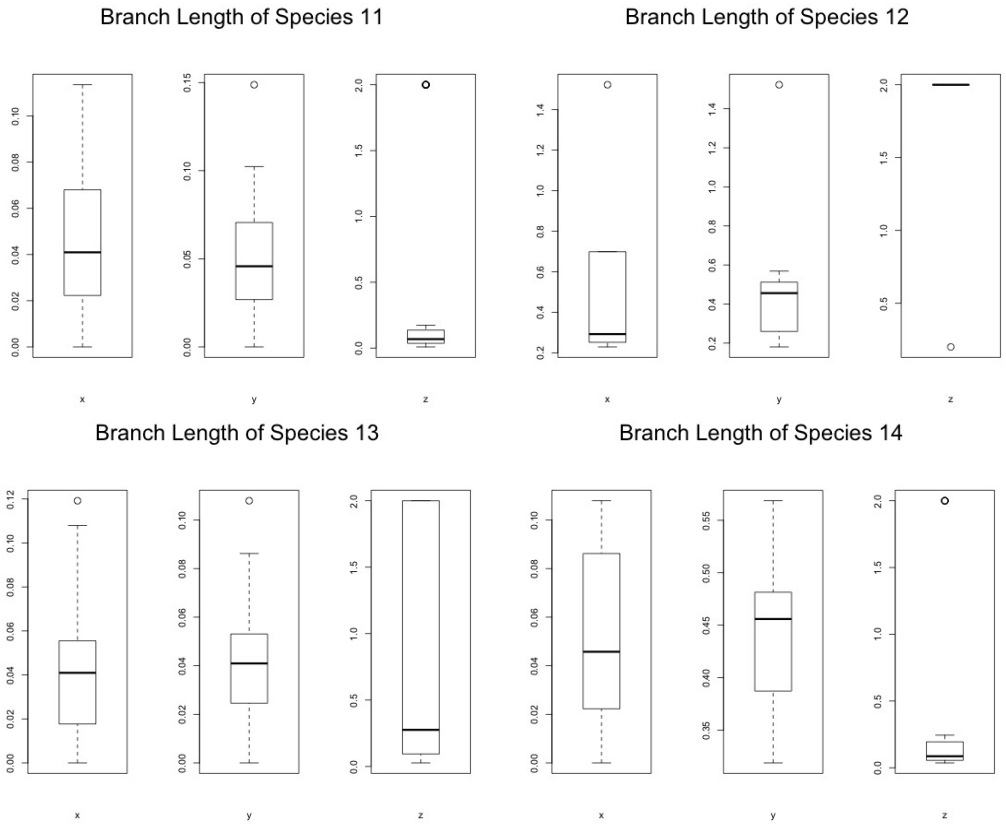

Figure 3.7: Box Plot of Pseudocaterpillar Species Branch Length

3.3 MLE and Bootstrapping Simulation for Five Taxa with DNA Sequences

This study uses the same method of the simulation for five taxa without DNA sequences to compute the MLE and bootstrapping for DNA sequences. However, the difference is that it is necessary to simulate the DNA sequences by using the Seq-Gen program and the PhyML program. After simulating the data from DNA sequences, it is necessary to follow the same steps that are used to compute the MLE and bootstrapping for five taxa without DNA sequences.

3.3.1 Result

Table 3.2: Average of Bootstrapping for Five Taxa With DNA Sequences

Type of Tree	Species Tree (x, y, z)	Prop. Correct	Average of Bootstrapping						
			1	2	3	4	5	6	7
Caterpillar	(0.1,0.1,1.0)	0.80	32.80	8.50	1.20	1.00	1.60	4.50	0.40
Balance	(0.1,0.1,1.0)	0.70	0.00	0.00	5.00	5.10	9.40	27.40	3.10
Pseudocaterpillar	(0.1,0.1,1.0)	100	1.10	1.10	0.60	2.60	42.40	1.70	0.50

Table 3.2 shows the proportion of correct matches of the MLE tree and the average bootstrap support for the MLE tree for species trees with DNA sequences. Both the caterpillar tree and the balanced tree in Table 3.2 with DNA sequences are less accurate than the same species trees without DNA sequences, as seen in Table 3.1, which matches the correct MLE tree. Also, the average bootstrap support the correct MLE tree for species trees with DNA sequences, which are less than the average of bootstrapping, and this supports the correct MLE tree for species trees without DNA sequences. This means that the MLE without the DNA sequences gives a better inference for the estimation of both species trees from the caterpillar

tree and balanced tree. Also, the result in Table 3.2 with a DNA sequence is higher than the result in Table 3.1, which lacks a DNA sequence. The pseudocaterpillar tree gives an accurate result from the DNA sequence. Finally, there is no significant correlation between the bootstrap support for the correct tree for MLE and the posterior probability based on both topologies and splits.

3.4 MLE of 5 taxa for equal branch length

This section investigates the MLE method with equal branch lengths for 5 taxa. The MLE method is simulated for caterpillar and balanced tree topologies with $x = y = z = 0, 0.1, 0.5, 1, 2$, and 3, and it uses the PhyloNet program (Than et al., 2008) to compute the MLE tree for rooted gene trees and unrooted gene trees, and retaining the best 100 proposed species trees, with 50 replicates for each combination of topology and branch lengths. All these computations have been done without DNA sequences. The details of how PhyloNet works in Chapter 4. The following subsection presents the results with equal branch lengths. Indicate that the simulation was done with both rooted and unrooted gene trees as input.

3.4.1 Results

Table 3.3 shows all possible clades in caterpillar shape for 5 taxa with equal branch length for rooted gene tree and for unrooted gene tree (abbreviated as GT in Table 3.3). The correct clades of the MLE tree for the caterpillar shape is AB, ABC, and ABCD. According to the results in the star tree (which has all branch lengths equal to 0) seen in Table 3.3, the branch length zero makes it difficult to match the caterpillar tree for both rooted gene trees and unrooted gene trees. But in the branch lengths 0.5, 1, 2, and 3, it is possible to see that the clades AB, ABC, and ABCD are

always obtained in the rooted gene tree. However, species trees with branch length 0.1 are also accurately inferred in the caterpillar clade shape for both rooted and unrooted gene trees. Clade AB is always obtained for branch length 0.5, 1, 2, and 3. However, clade ABC is always obtained for 0.5 and 1, but for branch length 2 and 3, is most highly obtained. The clade of ABCD is infrequently obtained with long branch lengths. However, clade DE is very frequently inferred with long branch lengths. Thus, the 5 taxa with long equal branch lengths are difficult to infer for the unrooted gene tree. These patterns are consistent with the results of ABC for equal branch lengths on caterpillar species trees.

Table 3.3: MLE for 5 Taxa Caterpillar Shape with Equal Branch Length and these results are out of 50 iterations.

	Species Tree (x, y, z)	AB	ABC	ABCD	DE	CDE	ABCE
Rooted GT	(0.0,0.0,0.0)	5	5	6	10	2	10
	(0.1,0.1,0.1)	42	46	47	1	0	2
	(0.5,0.5,0.5)	50	50	50	0	0	0
	(1.0,1.0,1.0)	50	50	50	0	0	0
	(2.0,2.0,2.0)	50	50	50	0	0	0
	(3.0,3.0,3.0)	50	50	50	0	0	0
Unrooted GT	(0.0,0.0,0.0)	3	8	10	4	1	11
	(0.1,0.1,0.1)	41	44	39	2	0	1
	(0.5,0.5,0.5)	50	50	26	6	0	18
	(1.0,1.0,1.0)	50	50	20	20	0	10
	(2.0,2.0,2.0)	48	38	4	46	10	0
	(3.0,3.0,3.0)	50	26	1	48	24	1

Table 3.4 shows all possible clades in the balanced shape for 5 taxa with equal branch lengths for rooted gene trees and for unrooted gene trees (abbreviated as GT in Table 3.4). The correct clades of the MLE tree for the balanced shape are AB, ABC, and DE. According to the results in Table 3.4 the star tree, which has branches of length 0 makes it difficult to infer the correct clade for matching the balanced tree for both rooted gene trees and unrooted gene trees. But in the branch

lengths 0.5, 1, 2, and 3, it is possible to see that the clades of AB, ABC, and DE is always accurately inferred in the rooted gene tree. However, species trees with branch length 0.1 are also accurately inferred for both rooted and unrooted gene trees. The clade AB is accurately inferred for all those branch lengths 0.5, 1, 2, and 3. The branch length 0.5 always matches all clades of the balanced tree. Long branch lengths 1, 2, and 3 lead to accurately inferring the clades AB, ABC, and DE, which leads to the inference that species of balanced shapes are easy to infer with long branch lengths. These results are also consistent with the results for the ABC method on five-taxon trees with equal branch lengths.

Table 3.4: MLE for 5 Taxa Balanced Shape with Equal Branch Length and these results are out of 50 iterations.

	Species Tree (x, y, z)	AB	ABC	DE	CDE	ABCD
Rooted GT	(0.0,0.0,0.0)	11	3	7	7	3
	(0.1,0.1,0.1)	43	49	46	1	3
	(0.5,0.5,0.5)	50	50	50	0	0
	(1.0,1.0,1.0)	50	50	50	0	0
	(2.0,2.0,2.0)	50	50	50	0	0
	(3.0,3.0,3.0)	50	50	50	0	0
Unrooted GT	(0.0,0.0,0.0)	7	8	5	4	8
	(0.1,0.1,0.1)	42	41	34	5	8
	(0.5,0.5,0.5)	50	50	50	0	0
	(1.0,1.0,1.0)	50	49	36	1	7
	(2.0,2.0,2.0)	48	30	47	20	1
	(3.0,3.0,3.0)	48	32	48	18	1

3.5 MLE for Eight Taxa

This study is done with MLE for five taxa, both with DNA sequences and without DNA sequences. This study the attempts to apply MLE to more numbers of taxa,

which in this study consists of 8-taxa. This section has two subsections: The first is the method that this study needs in order to do the computation, and another subsection is the result of the computation.

3.5.1 Method of Simulated the MLE and Bootstrapping for Eight Taxa

To compute the MLE and bootstrapping, we simulated the gene trees by using the Hybrid-Lambda program from the observed data (Zhu et al., 2015). After that, it is necessary to simulate the observed data which is used as the input data for the PhyloNet program to compute the MLE (Than et al., 2008).[1] The PhyloNet program needs three files to run. The PhyloNet program works with both rooted and unrooted gene trees.

3.5.2 Result

Table 3.5 shows that the percentage of the bootstrapping supports the MLE method for the 8-taxa caterpillar shape. The consensus tree matches the MLE tree for caterpillar shape. Moreover, the MLE tree obtains the clade of {ABCDE} by 100% in all iterations. Also, Figure 3.8 shows the caterpillar tree with the percentage of each clade and Figure 3.9 also shows the percentage of bootstrap support in each clade, which supports the MLE tree for the caterpillar tree.

Table 3.6 shows that the percentage of the bootstrapping supports the MLE for the 8-taxa balanced shape. The consensus tree matches the MLE tree for balanced shape. Also, Figure 3.10 shows the balanced tree with the percentage in each clade

[1]This simulation needs to use the Java program since the program PhyloNet needs Java in order to run.

Table 3.5: The percentages of the MLE Tree and the Bootstrap with a Caterpillar Shape for 8 taxa.

Type of Method	AB	ABC	ABCD	ABCDE	ABCDEF	ABCDEFG
MLE	0.28	0.40	0.56	100	0.36	0.74
Average Bootstrapping	0.350	0.401	0.602	0.997	0.346	0.611

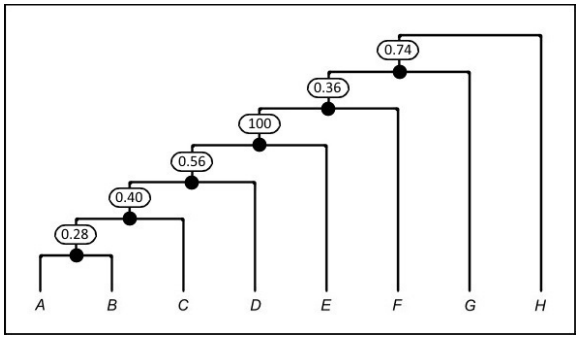

Figure 3.8: MLE for caterpillar tree of 8-Taxa

and Figure 3.11 also shows the bootstrap percentage in each clade, which supports the MLE tree for the balanced tree.

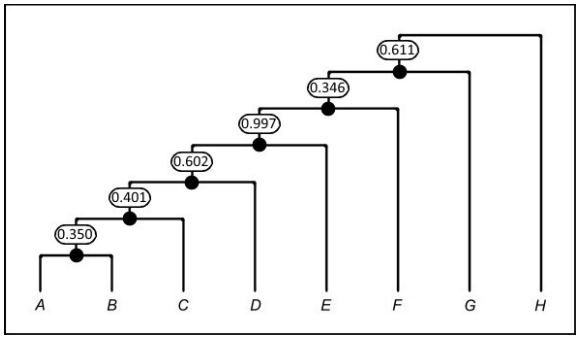

Figure 3.9: Bootstrapping supporting MLE tree for caterpillar shape of 8-taxa

Table 3.6: The percentages of the MLE Tree and the Bootstrap with a Balanced Shape for 8 taxa.

Type of Method	AB	CD	ABCD	EF	GH	EFGH
MLE	0.90	0.54	0.96	0.40	0.64	0.98
Average Bootstrapping	0.88	0.55	0.95	0.45	0.51	0.98

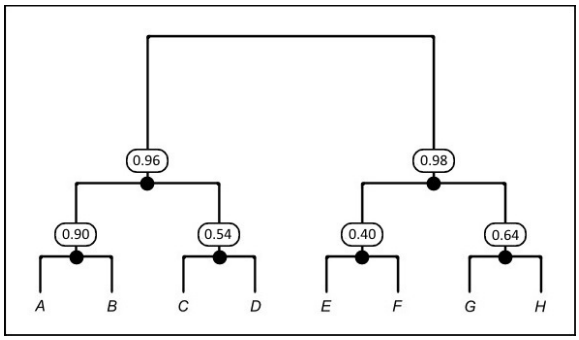

Figure 3.10: MLE for balanced tree of 8-taxa

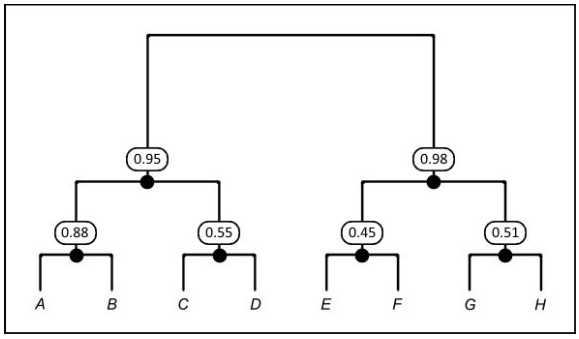

Figure 3.11: Bootstrapping supporting MLE tree for balance shape of 8-Taxa

Chapter 4

Inferring Species Trees From Rooted vs Unrooted Gene Trees

This chapter compares using unrooted gene trees to infer a rooted phylogeny to using rooted gene trees when there is an outgroup and molecular clock (with DNA sequences and without DNA sequences). Comparisons of unrooted versus rooted gene tree methods are made using ML only.

This approach both assumes that at least one taxon, the outgroup, is outside of the ingroup, and that the root of the ingroup is a branch at which the ingroup and the outgroup connect. The outgroup method indicates the root's location in the analysis. To perform midpoint rooting, it is first necessary to estimate the ML tree – this must be done with no recourse to taxa from the outgroup. It is then necessary for the root to be placed in the middle of two terminal taxa with the longest path between them. To satisfy the molecular clock (and end up with rooted trees), it is necessary to constrain the ML and Bayesian clock analyses so that the sum of branch lengths from root to tip is constant – this also must be done with no recourse to taxa from the outgroup (Boykin et al., 2010). This chapter contains two sections: One

explains the method of simulation, and the other shows the results of the study.

4.1 Method of simulation

There exist four conditions under which species trees are inferred. The first type consists of rooted, known gene trees. The second type consists of rooted gene trees estimated from DNA sequences. The third type consists of unrooted, known gene trees. Finally, the fourth and last type consists of unrooted gene trees estimated from DNA sequences.

This study requires seven steps in order to do the simulation. The first step uses the library from the R-package, which is TreeSim (Stadler, 2014). The TreeSim library simulates the species tree. To simulate the species tree by using the library TreeSim in the present study, it is necessary to have the following important information. It is first necessary to know the number of taxa, which, in the case of this study, are of two different varieties, which are 5 taxa and 8 taxa. Then, it is necessary to determine how many trees to simulate. It is also necessary to specify the value of λ (birth rate). In this study, λ has five difference values, which are 0.1, 0.25, 0.5, 0.75, and 1.0. The value of μ/λ (turnover) is the final information needed to simulate the species tree. There are four values of μ/λ, which are 0.0, 0.25, 0.5, and 0.75.

The second step is to add the outgroup to the species tree. The present work uses R code to do it. Two versions of the species tree are saved, which makes one with the outgroup to compare with the result from the PhyloNet program for rooted gene trees and the other version without the outgroup to compare with the results from the PhyloNet program for unrooted gene trees. The third step is to run the ultrametric program, which makes the simulated tree from the first step into a molecular clock tree with the outgroup. The fourth step is to use the molecular clock species tree as

an input to the Hybrid-Lambda program to simulate the gene trees. The fifth step is the longest step since the present work needs to run the PhyloNet program.

The sixth step is to simulate DNA sequences by using the Seq-Gen program and the PhyML program. The Seq-Gen program uses the gene tree that is acquired from Hybrid-Lambda as input to the program. The PhyML program uses the output data from Seq-Gen as input. The present work runs an R code to make the output data from PhyML as a rooted tree and save it as data simulated from DNA sequences. The final step, which is the seventh step, is very similar to the fourth step in the present study. The outgroup is removed after the gene trees are obtained from the hybrid-lambda program to compute the MLE from the unrooted gene trees without DNA by using the PhyloNet program. Also, the outgroup needs to be removed from the DNA sequences of the PhyML program to compute the MLE of the unrooted gene trees with DNA by using the PhyloNet program. Figure 4.1 shows the steps for building the simulation code and how they compare with the results of the present study. All R codes and the script are included in Appendix A.4.

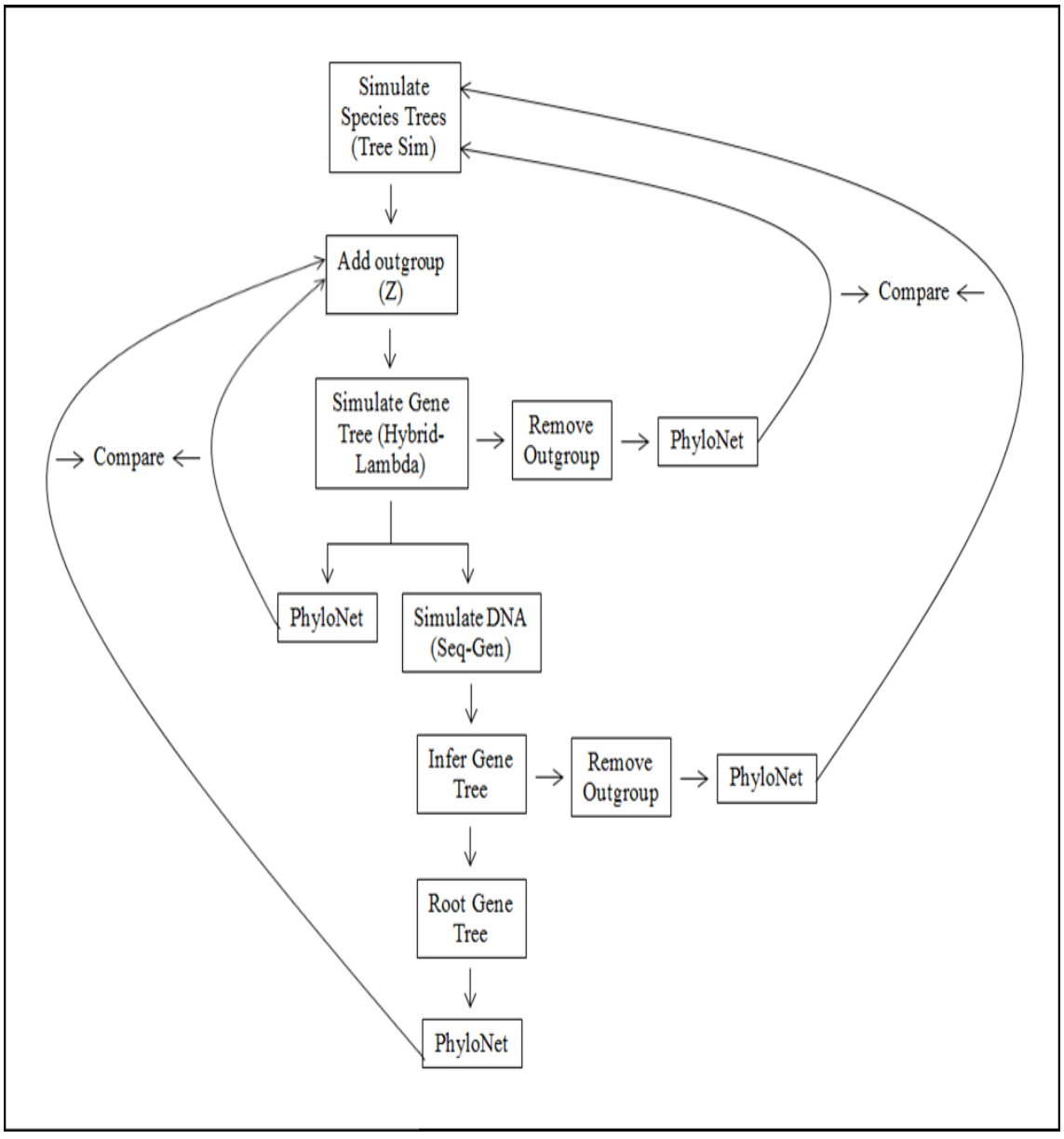

Figure 4.1: Diagram of the Simulation Method.

4.2 Result

Table 4.1: Rooted Gene Trees vs Unrooted Gene Trees for Five Taxa

λ	μ	RGT without DNA	RGT with DNA	URGT without DNA	URGT with DNA
0.10	0.00	0.16	0.24	1.6	1.52
	0.25	0.24	0.2	1.72	1.68
	0.50	0.12	0.12	1.8	1.68
	0.75	0.12	0.12	1.88	1.68
0.25	0.00	0.40	0.44	1.32	1.20
	0.25	0.16	0.36	1.56	1.32
	0.50	0.24	0.48	1.36	1.52
	0.75	0.36	0.40	1.48	1.40
0.50	0.00	0.64	0.67	1.24	1.32
	0.25	0.72	0.56	1.12	1.08
	0.50	0.24	0.52	1.00	1.12
	0.75	0.28	0.44	1.16	1.24
0.75	0.00	1.04	1.00	1.12	1.20
	0.25	0.72	0.72	0.84	1.32
	0.50	0.64	0.96	1.28	1.20
	0.75	0.48	0.60	0.88	1.36
1.00	0.00	1.00	1.04	1.32	1.16
	0.25	0.40	0.76	1.08	0.96
	0.50	0.96	0.92	0.96	1.20
	0.75	0.64	0.80	1.04	1.28

Table 4.1 shows the average of the tree distance for 5 taxa between the simulated species trees with the outgroup and without the outgroup; then it is necessary to compute the MLE trees for rooted gene trees with DNA and without DNA, and it also has been done for unrooted gene trees with DNA and without DNA. When the value of λ increases, inferring the species tree from rooted gene trees becomes more difficult, but in most of cases, when the values of μ increase, it reduces the difficulty of inferring the species tree from rooted gene trees. In all values of λ and μ, the averages of rooted gene trees with DNA and without DNA are less than the average

of unrooted gene trees with DNA and without DNA, which means the rooted gene trees with DNA and without DNA are more accurate. Since the maximum distance between trees is given by this formula $2n - 4$, the maximum number for missing 5-taxon nodes is 6. According to all results in Table 4.1 and also between all four types of comparisons of trees, it is possible to conclude that using rooted gene trees is more accurate than using unrooted gene trees.

Table 4.2: Rooted Gene Trees vs Unrooted Gene Trees for 8 Taxa

λ	μ	RGT without DNA	RGT with DNA	URGT without DNA	URGT with DNA
0.10	0.00	0.32	0.52	1.60	3.04
	0.25	0.12	0.28	1.68	2.68
	0.50	0.20	0.40	1.40	2.52
	0.75	0.08	0.40	1.68	3.00
0.25	0.00	0.64	1.16	1.76	2.12
	0.25	0.56	1.24	1.52	2.40
	0.50	0.52	0.72	1.28	2.12
	0.75	0.52	0.56	2.24	2.92
0.50	0.00	2.28	2.76	3.08	2.84
	0.25	1.48	2.00	2.44	2.72
	0.50	2.04	1.88	2.28	3.12
	0.75	1.44	2.20	2.32	3.2
0.75	0.00	2.96	3.52	2.92	3.32
	0.25	3.04	3.64	2.24	3.24
	0.50	2.24	2.48	2.64	2.76
	0.75	2.00	2.88	2.76	3.04
1.00	0.00	3.40	4.04	2.96	3.88
	0.25	2.52	3.60	2.84	3.00
	0.50	3.16	3.32	2.88	3.48
	0.75	2.52	2.76	2.48	3.04

Table 4.2 shows the average tree distance for 8 taxa between the simulated species trees with the outgroup and without the outgroup; then it is necessary to compute the MLE trees for rooted gene trees with DNA and without DNA, and it is also necessary for unrooted gene trees with DNA and without DNA. When the

value of λ increases, inferring species trees from rooted gene trees becomes more difficult, but in most of cases, when the values of μ increase, it reduces the difficulty of inferring the species tree from rooted gene trees. The maximum distance between 8-taxon trees is 12. In most of the values of λ and μ, the averages of rooted gene trees with DNA and without DNA are less than the average of unrooted gene trees with DNA and without DNA, which means the rooted gene trees with DNA and without DNA are more accurate. However, when λ becomes bigger, according to Table 4.2 for $\lambda = 0.75$ and 1, with increasing the value of μ, then most of the unrooted gene trees without DNA do better than the rooted gene trees. According to all results in Table 4.2 and also between all four types of comparisons of trees, it is possible to conclude that the rooted gene tree is more accurate in the first three values of λ. But in the last two value of λ, it is possible to say that the unrooted gene tree without DNA appears to be slightly more accurate.

Figure 4.2 and Figure 4.3 show the comparison between rooted gene trees and unrooted gene trees for cases that have the gene trees with and without DNA for both 5 taxa and 8 taxa. On the horizontal axis are the values of λ, and in the vertical axis is the average of the tree distance to the species tree from both rooted gene trees and unrooted gene trees with and without DNA, where the μ is constant in each subfigure. Those figures show how the λ changes when μ is increased. Figure 4.2 and Figure 4.3 display that there are not many differences between the rooted gene trees and unrooted gene trees in both cases with and without DNA. However, Figure 4.2 shows that when the λ is increased the variation of differences decreases with all values of μ. Moreover, when we increase the numbers of taxa from 5 taxa to 8 taxa, Figure 4.3 shows that unrooted gene trees become more accurate in all values of μ when λ is big. However, the rooted gene tree without DNA makes good estimates for 5 taxa in all values of λ and μ, as seen in Figure 4.2. For 8-taxa, the rooted gene tree without DNA makes a better estimate with small values of λ, but when λ becomes bigger, the unrooted gene tree without DNA does better, as seen in

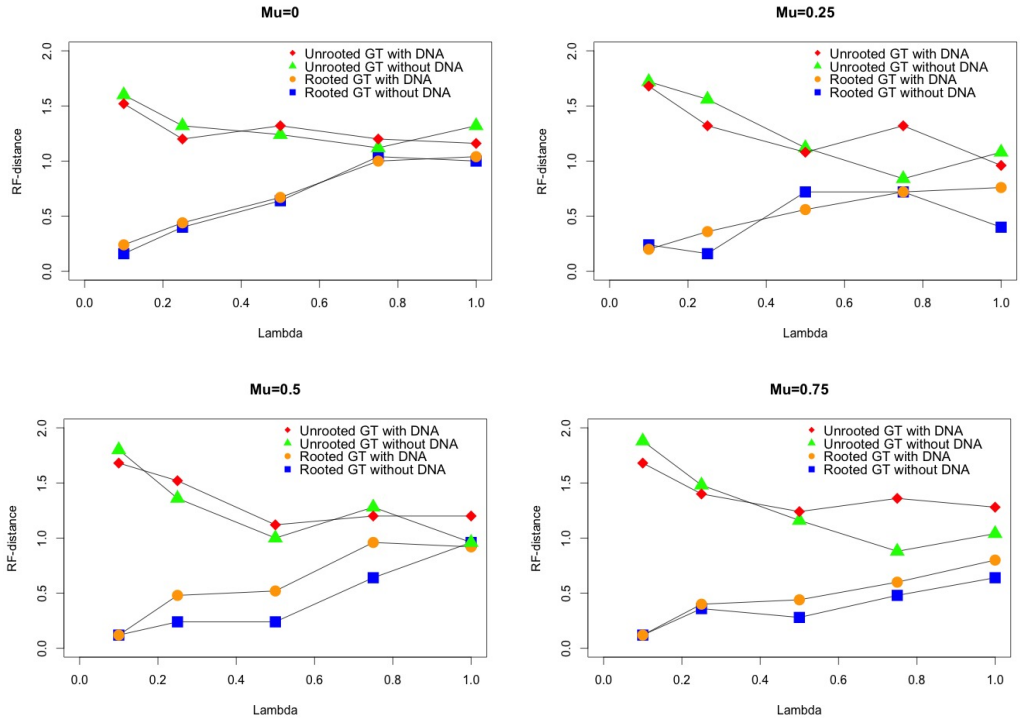

Figure 4.2: Rooted GT vs Unrooted GT for 5 Taxa

Figure 4.3. Since unrooted gene trees make better estimates by increasing the value of λ, when we increase the number of taxa, it may end up giving a good estimate for a big number of taxa.

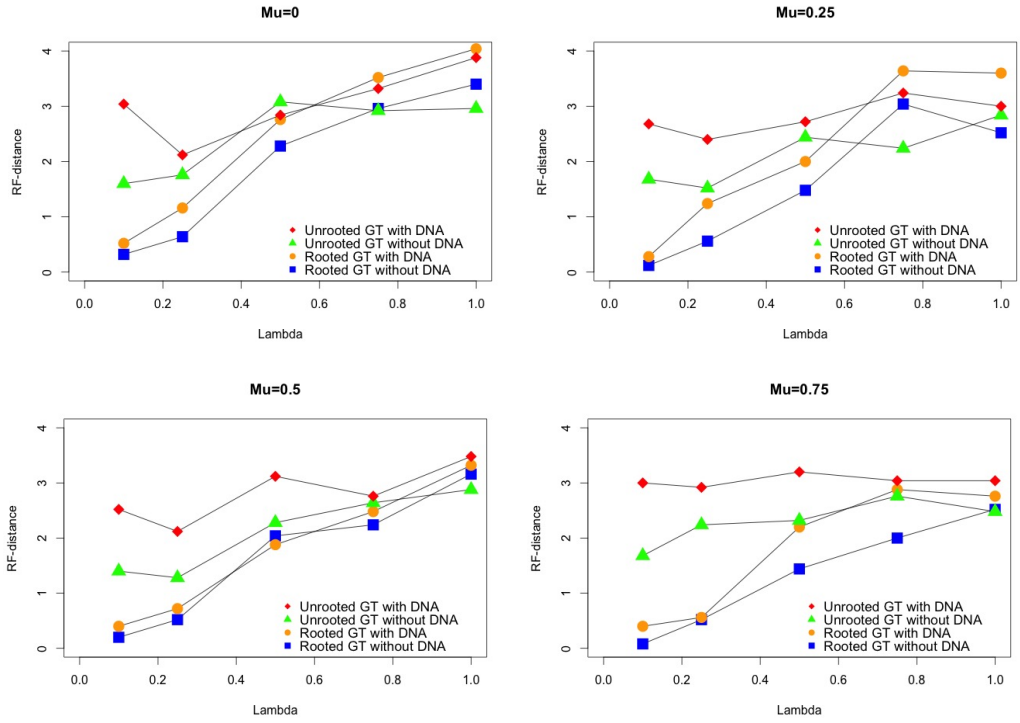

Figure 4.3: Rooted GT vs Unrooted GT for 8 Taxa

Chapter 5

Empirical Study

This chapter concentrates on an empirical example, in order to see how the method of the present simulation study, in both Chapter 2 and Chapter 3, works with empirical data. The empirical data from the Xi et al. (2014) was chosen to apply methods of the present study in an attempt to see which hypotheses of their study are supported when Amborella clusters with Nuphar or when Nuphar clusters with the all species in question.

The present work thus seeks to apply the ABC method, used elsewhere in the present work, to their two hypotheses with the aim of determining which one works best in the present framework. The present study also seeks to apply the MLE method and bootstrapping, used elsewhere in the present work, to the two hypotheses of their work with the aim of determining which one works best in the present framework.

5.1 Method

The present study incorporates data for 310 nuclear genes found in 45 seed plants, which were analyzed by Xi et al. (2014). They argue that coalescent methods produce better results than concatenation methods for fast-evolving nucleotide sites. A subset for eight taxa is taken from this data with the aim of finding eight common tip labels between all those 310 nuclear genes, but there were considerable amounts of missing taxa for many loci. This ends up giving 224 nuclear genes for eight taxa, as shown in Figure 5.1.

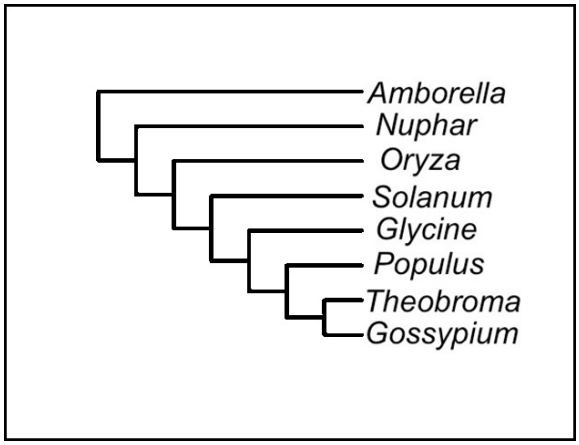

Figure 5.1: A subset species tree from a species tree of the Xi et al. (2014) study for 8-taxa

To apply the ABC method (Algorithm 3) in this empirical data for 8 taxa, it is necessary to use the following steps. The first step is to run the empirical data with a split program in order to compute the splits of the empirical data and compare it with the splits of the simulation data from the prior. The second step is to generate the data to use it as input for the Hybrid-Lambda program from the prior, which uses the hypotheses of Xi et al. (2014) as priors - their first hypothesis assumes that Amborella clusters with Nuphar, and their second hypothesis assumes that Nuphar clusters with all other as shown in Figure 5.2. The third step is to modify the trees

generated from the prior to be ultrametric. The fourth step is to use the species tree from the prior as input for the Hybrid-Lambda program to simulate gene trees. The fifth step is to run the split program again with the simulated gene trees to compute the splits of simulation data. The final step is to run some R code to compute the distance between the split frequencies of the empirical data and the split frequencies of the simulated data and also to use the consensus program to compute the posterior probabilities.

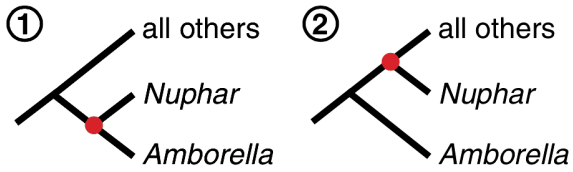

Figure 5.2: Two hypotheses by Xi et al. (2014) used as the prior for the present study (p. 922).

The MLE for the empirical data has been done by using the PhyloNet program to compute the MLE tree. Moreover, the bootstrapping for the MLE tree has also been done by using the PhyloNet program. The previous section provides an explanation of how the PhyloNet program works. The empirical study does not need to simulate the data since the empirical data is directly used as an input to create the PhyML-middle file to run the PhyloNet program.

For five taxa case, the subset from this data is also taken with the aim of finding five common tip labels between all of the 310 nuclear genes. This ends up giving 275 nuclear genes for five taxa, as shown in Figure 5.3.

After getting this empirical data for five taxa, it is possible to apply the same two algorithms (1 and 2) used for the ABC method as the five-taxon cases in chapter 3 and 4. The MLE approach is the same as in chapter 4. In these empirical five taxa, the same code used to compute regular five taxa without DNA are used, but

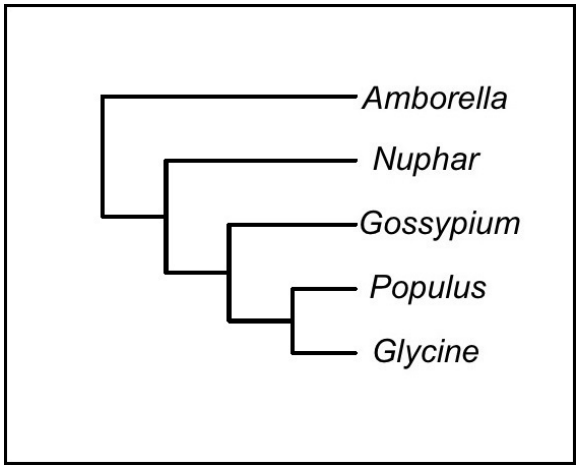

Figure 5.3: A subset of a species tree from the Xi et al. (2014) study for 5-taxa

the difference involves counting the topology of the empirical five taxa directly from the empirical data, which is called the observed topology counts observation. The vector program is also used to compute the topology counts of the simulated data that is obtained from the prior, which is the prior dependent of the Xi et al. (2014) hypothesis, as shown in Figure 5.2. The next step is to compute the distance from the observed topology frequencies and the simulated topology frequences, which leads to computing the ABC method.

The degree to which the MLE tree and bootstrapping method support the MLE tree was calculated in two different ways. For the first way, the same code computing the MLE and bootstrapping for five taxa with DNA and without DNA was used. For the second way, the PhyloNet program was used to compute the MLE and bootstrapping, and the consensus program was then used.

5.2 Results

Figure 5.4 shows both tree shapes in both hypotheses for five taxa with the tip-label name of the genes. The posterior probability of the topology of the empirical data for five taxa supports the Amborella clustering with the Nuphar by 86%. In contrast, 14% of the posterior probability of the topology supports the Nuphar clustering with the others genes. For the posterior probability based on splits, 86% of the results support Amborella clustering with Nuphar, and 14% of the results support Nuphar clustering with other genes. The MLE and bootstrapping from the present study's code without using the PhyloNet program show a tree that has the Amborella cluster with the Nuphar, which is also supported by bootstrapping in 100% of the results. From all of the results obtained from study, the five taxa empirical data highly support that the Amborella clusters with the Nuphar.

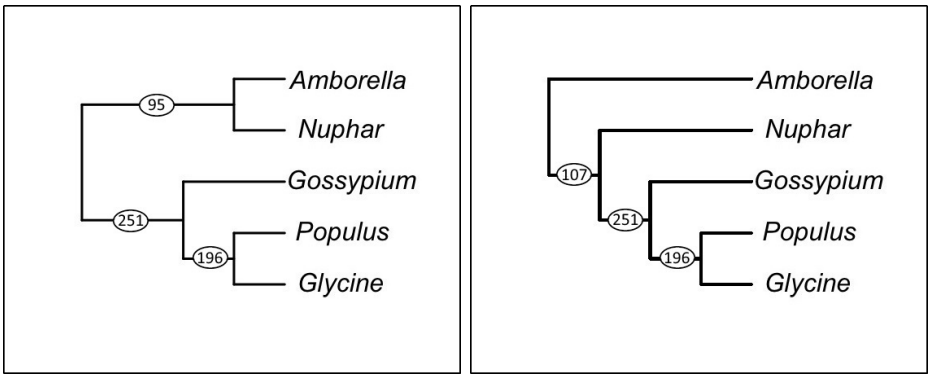

Figure 5.4: Two hypotheses by Xi et al. (2014) used as the prior for 5-taxa in the present study (p. 922). Numbers on branches represent the number of trees in the input with the given clade.

Figure 5.4 shows the 5-taxa of the empirical data from the consensus program, which supports the hypothesis of the Xi et al. (2014). Figure 5.5 shows the 8-taxa of the empirical data from the consensus program which supports the hypothesis of Xi et al. (2014), which is that the Nuphar clusters with other genera and the Nuphar

culster with Amborella, but after computing the ABC method and the MLE method with support from the bootstrap method, the results of the ABC method support the clade of Amborella clustering with the Nuphar by 59%, as shown in Figure 5.6.

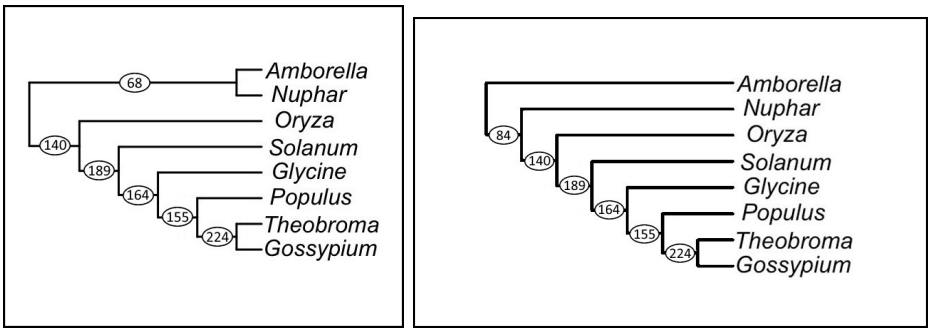

Figure 5.5: Empirical Shape from consistent program for 8 Taxa

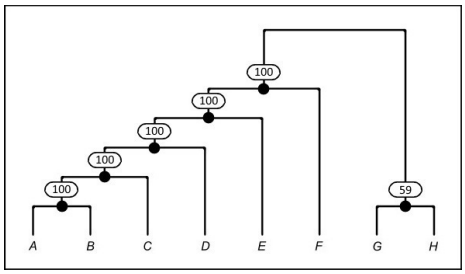

Figure 5.6: Empirical Tree for 8 Taxa by ABC Method

The MLE method shows greater support for the Amborella and Nuphar clade than for the Amborella-as-outgroup hypothesis; however, the support is not overwhelming, with bootstrapping for this hypothesis being either 80% or 62% for the eight-taxon and five-taxon analyses, respectively. This lends some support, but not overwhelming support, to the hypothesis of Xi et al. (2014) as opposed to the hypothesis supported by concatenation, using a subset of their gene trees. This is perhaps not surprising since Alanzi and Degnan (forthcoming) use the same gene tree topologies that lead to the conclusion of Xi et al. (2014). Alanzi and Degnan (forthcoming) also use a coalescent method, although it is quite different from the

methods used in the study of Xi et al. (2014).

Simmons and Gatesy (2015) argue that the anomaly zone, in which the most likely gene tree has a different topology from that of the species tree, does not apply to the root of the rooting of the angiosperms because there is only one short branch leading to the Amborella-Nuphar clade, whereas the anomaly zone typically occurs when there are two consecutive branches on a path from the root to the tips (Degnan, 2013; Rosenberg, 2013). Although this is typical of species trees in the anomaly zone, a species tree with only one short branch can still be in the anomaly zone if the less basal branch can be indefinitely long if the more basal branch is sufficiently short (Degnan and Rosenberg, 2006). However, for this example, there is another reason for thinking that the anomaly zone is not a factor, which is that caterpillar gene tree shapes, which the Amborella-only outgroup hypothesis implies for the subset of species that Alanzi and Degnan (forthcoming) analyzed, cannot be anomalous gene trees (Degnan and Rhodes, 2015).

Another consideration is the clade support for Amborella as an outgroup. Simmons and Gatesy (2015) point out that the Amborella-only outgroup occurs more often than any conflicting relationship. Under the multispecies coalescent model, any clade that occurs with more than 1/3 probability in the true gene trees is guaranteed to be in the species tree (Allman et al., 2011a). Using the 275-locus five-taxon dataset, the proportion of loci for which Amborella is the only out- group is 107/275 = 38.9% as shown in Figure 5.4, thus suggesting some evidence using coalescent considerations that the Amborella- outgroup hypothesis is correct. The eight-taxon data set with more trees has weaker evidence, with 84/224=37.5% as shown in Figure 5.5 of trees supporting the Amborella- hypothesis. Although the proportion of trees supporting the Amborella-only outgroup hypothesis is larger than for trees with the Amborella-Nuphar clade, these differences in proportion are also not significant. Simmons and Gatesy (2015) also criticize the data analysis that lead to the 310 gene

trees of Xi et al. (2014) particularly in terms of rooting, but also in terms of the alignments used to reconstruct the gene trees. Simmons and Gatesy (2015) prefer the summary coalescent analysis of Wickett et al. (2014), which used the unrooted method ASTRAL (Mirarab et al., 2014) and found support for the Amborella-only outgroup hypothesis. Our analysis is not intended to take sides in this debate; instead, Alanzi and Degnan (forthcoming) use the Xi et al. (2014) data to illustrate how the ABC approach and the MLE approach can be used to estimate a root from unrooted gene trees, and find that this data set has the interesting property that a set of gene trees can lead to a naive consensus method returning a caterpillar tree while other methods can lead to a non-caterpillar estimate of the species tree.

Chapter 6

Conclusion and Discussion

Table 6.1: Example to illustrate that split counts are not sufficient statistics

Dataset	tree	splits		
1	t_1	ABC	DE, BC	ADE
1	t_2	ABD	CE, AB	CDE
1	t_3	BCD	AE, BD	ACE
2	t_4	ABC	DE, AB	CDE
2	t_5	ABD	CE, BD	ACE
2	t_6	BCD	AE, BC	ADE

Although the ABC approach in the present study used counts of split as a summary statistic, it is notable that this statistic is not sufficient statistic. There is not always a guarantee that the estimated and true posterior distributions will converge with each other; for example, when the statistical summaries are given for the ABC method results, they are sometimes insufficient for this purpose (Marjoram and Tavaré, 2006; Csilléry et al., 2010). This is typical of analyses using the ABC method (Aeschbacher et al., 2012). In our case, split counts are not sufficient statistics, although they do identify the species tree (Allman et al., 2016), meaning that knowing the probabilities of the splits allows the inference of a unique species tree. In particular, for a summary statistic T to be a sufficient statistic, it should

be the case that given two data sets x, and y, if $T(x) = T(y)$, then any inference about the parameter should be the same for the two data sets (Casella and Berger, 2002). However, an example of two data sets each with three trees illustrates that split counts can be the same for two data sets with different input trees as shown in Table 6.1.

For this example, the split counts for the two datasets are identical, but the original trees are different, and the likelihoods for the two data sets are different based on using unrooted gene tree probabilities listed in Allman et al. (2011b). Although split counts are not, therefore, sufficient statistics, they still identify the species tree in the sense that two distinct species trees necessarily have different probabilities of splits in the gene trees (Allman et al., 2016). The five-taxon examples suggested that split counts did essentially as well as topology counts, suggesting that there was very little loss of information in using splits instead of topologies, in spite of the lack of sufficiency. This study also supports Allman et al. (2011b), who mention it in their paper by knowing the unrooted topology from which all desired information about the rooted tree can be obtained.

In cases where the ST-ABC method drew from the set of retained trees and used the most frequently occurring one among them as an estimate for species trees, the method faced performance difficulties in all 14 of the cases examined in the simulation study involving 5 taxa. The simulation needed a long time to finish all shapes of taxa, which was difficult since time is limited. For example, this is a sample calculation of time that this study needed to finish: 14×50×2=1400 days, which means that 4 CPU years are needed for doing the ABC for five taxa without DNA. Another sample time is 14×50×1=700 days, which means that 2 CPU years are required for doing the MLE with bootstrapping for five taxa without DNA. Moreover, the five taxa with DNA require 2×400×3=2400 days, which is equal to 7 CPU years to finish. That is also for the ABC of eight taxa 2×400×2=1600 days,

which is more than 4 CPU years.

This study concludes that five-taxon species trees are difficult to infer since, doing this requires a greater number of loci. Since the correlation between the average of the posterior probability for topology and splits is very strong this suggests that all the information about the topology can be known by studying the split of species tree for five taxa without DNA. There is not much difference in the results using topologies and split but sometimes the split makes much better results. The consensus program can be used program for summarizing the ABC posterior distribution. Increasing sample size helps to get a lot of accuracy with the estimate. Sequence data is not needed for the ST-ABC method since gene trees are its only source of data. However, it is assumed that the gene trees in the sample are all known with certainty. Despite shortcomings with the ST-ABC method, larger sample sizes may increase the accuracy and decrease the variability of the estimates. Whether or not the branch lengths referred to are short ($e.g; 0.1$ coalescent units) or long ($e.g; 1$ coalescent units), accurately estimating a species tree topology can sometimes be done with a sample size of $N = 100$ loci or more than 100 loci.

The ABC method could be used with a flat prior or a more informative prior. This study used a prior that was uniform for topologies but assumed that a particular unrooted species tree was known. If the unrooted species tree also had uncertainty, then this could be reflected by making the prior include more rooted species trees. Another possibility is to consider that under typical birth-death processes, some unrooted trees are more likely than others when there are 6 ore more taxa (Steel, 2012), and the prior could be based on this rather than making each labeled topology equally likely in the prior.

If midpoint rooting gives one tree and an out-group gives another a prior could be fifty-fifty for these two species trees. The midpoint and out-group give a different prior, which makes the ABC more efficient to compute, as well as easier to deal

with, rather than going by all uniform priors of the species trees (Boykin et al., 2010). For example, there is an empirical question about how many rooted trees in the taxa are required to account for the 90% making the posterior probability (Boykin et al., 2010). In comparing the two, the results for the five DNA taxa are less accurate than the results for the five regular taxa at counting the average posterior probability. Since it is difficult to estimate the root location with certainty, a possible application is to instead rule out implausible locations for the root, and in this case a credible region for the root might be desirable. In the five-taxon cases, 90% credibility regions tended to have about 3-5 trees, meaning that about half of the possible root locations had very low posterior probability.

The conclusion is, thus, that the ABC and MLE methods used in Chapter 2 and Chapter 3 yield the same results. It is difficult to infer the species tree with caterpillar shape for unrooted gene trees with long branch length. However, the result of the balanced shape with equal branch length, which is computed by the ABC method, agrees with the MLE method. This assumes that it is easy to infer the species tree with long branch length. Both methods agree that it is difficult to infer the star tree.

The MLE method makes somewhat more accurate inferences than the ABC method does because the MLE shows better in about 7 out of 13 cases for topologies and 8 out of 13 cases for splits as seen in Figure 6.1. Figure 6.1 top shows the relation for the proportion by which both the MLE method and the ABC method match the correct tree by using topology counts. Figure 6.1 bottom displays the relation for the proportion by which both the MLE method and the ABC method match the correct tree by using split counts. In the $x = y$ line graph, the points above the line indicate that the proportion of matching a correct MLE tree was higher than using topology counts. Similarly, the points above the line indicate that the proportion of matching a correct MLE tree was higher than using split counts as the summary statistic.

The proportion for the correct tree tended to be slightly lower when using the ABC method with splits or topologies, but highly correlates with the proportion using the MLE method. PhyloNet is a very easy way to compare unrooted gene trees and rooted gene trees. It is necessary to infer the phylogenetic tree by using MLE to figure out which is the maximum likelihood tree and then to use the bootstrapping method to measure the support of the MLE tree. In most of the computation, the MLE tree has support by the bootstrapping even though the MLE does not match the correct tree.

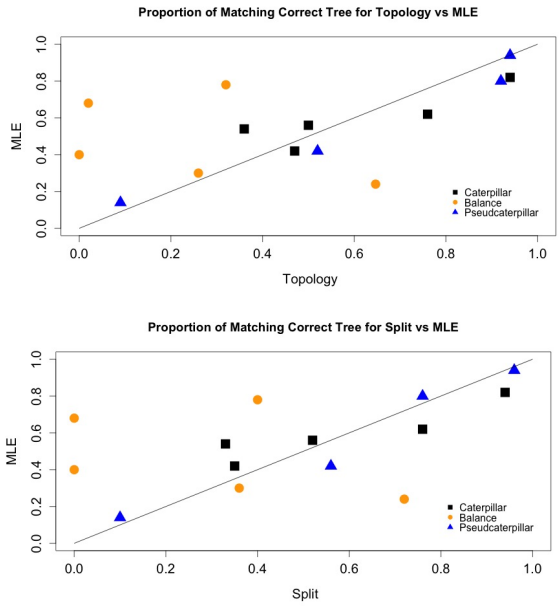

Figure 6.1: Proportion of the MLE method and the ABC method match the correct tree

The ABC method, MLE method, and bootstrap method in the empirical study for both cases of 8-taxa and 5-taxa support that Amborella clusters with Nuphar by more than 2/3 of the posterior probability. In the comparison between the posterior probability using topologies and the posterior probability using splits, according to the empirical data, it is found that the posterior probability using splits obtained

the same probability.

According to the results from the empirical data, the ABC method and MLE method weakly support the conclusion of Xi et al. (2014) that Amborella and Nuphar cluster together based on their gene tree topologies, but using very different methodology (although theirs is also coalescent-based). Alanzi and Degnan (forthcoming) note that their gene trees have been criticized, in particular for having problems with how they are rooted. However, our method uses only unrooted gene tree topologies and still found some support for their estimated species tree. Alanzi and Degnan (forthcoming) have used this data set not to make a definitive claim about the rooting of the angiosperms, but rather to illustrate how the ABC method and MLE method could be used for this type of problem.

Appendix A

Script and R code

A.1

1. Code for computing a prior, which is "priorst.r".

$library(ape)$

$tree <- -sample(1:7, 1)$

$prior.tree <- -paste(\text{“}prior\text{”}, tree, sep = \text{“”})$

$prior <- -read.tree(prior.tree)$

$prior\$edge.length = rexp(length(prior\$edge.length), rate = 1)$

$write.tree(prior, file = \text{“}tempst.1\text{”})$

2. Script for computing the ABC method, MLE, and Bootstrapping, which is "scriptABC".

cd name direction # This is used to open the direction and save the output in this direction

$for((h = 1; h <= 1; h++))$.

do

rm -f vector_ts5a.obs # This command is used to delete this file.

rm -f vector_ts5a.sim

rm -f st_sim

rm -f Distance.output

./hybrid-Lambda -spcu ABCDE -num 100 -seed 130 # This command is used to run the hybrid−Lambda program to generate the gene from the observed data.

sed 's/_1//g' OUT_coal_unit > data-ts5a

$for((i = 1; i <= 100; i++))$

do

head −$i data-ts5a | tail −1 > gt

./vector gt 0 >> vector_ts5a.obs # This command is used to run the vector program to compute the topology from the observe data.

done

$for((j = 1; j <= 50000; j++))$

do

rm −f vector_ts5a.sim

R CMD BATCH priorst.r # This command is used to run the R code to compute the prior.

./ultrametric tempst.1 > ts5a.new # This command is used to run the ultrametric program to make the prior data into a molecular clock.

cat ts5a.new >> st_sim

echo $h $j >> log-output-number

./hybrid−Lambda −spcu ts5a.new −num 100 −seed 1301 # This command is used to the run hybrid−Lambda program to generate the gene from data that is acquired from the prior.

sed 's/_1//g' OUT_coal_unit > data−ts5a.new

$for((k = 1; k <= 100; k++))$

do

head -$k data-ts5a.new | tail -1 > gt.1
./vector gt.1 0 >> vector_ts5a.sim # This command is used to run the vector program to compute the topology from the simulated data.
done
R CMD BATCH compare.r.2 # This command is used to run the R code to compute the topology for the five taxa and also to compute the split of the five taxa.
done
R CMD BATCH post.boot.r # This command is used to run the R code to compute the posterior probability of the topology, the posterior probability of the split, the MLE, and also bootstrapping.
done

3. R Code for computing the topology of the five taxa and the splits of the five taxa, which is "compare.r.2".

x.topo.obs $<-$ read.table("vector_ts5a.obs")
n.topo.obs$<-$ 1:15
$n.topo.obs[1] <- sum(x.topo.obs ==$ "2_3_4_4_3_4_4_3_3_2")
$n.topo.obs[2] <- sum(x.topo.obs ==$ "2_4_3_4_4_3_4_3_2_3")
$n.topo.obs[3] <- sum(x.topo.obs ==$ "2_4_4_3_4_4_3_2_3_3")
$n.topo.obs[4] <- sum(x.topo.obs ==$ "3_2_4_4_3_3_3_4_4_2")
$n.topo.obs[5] <- sum(x.topo.obs ==$ "4_2_3_4_4_3_2_3_4_3")
$n.topo.obs[6] <- sum(x.topo.obs ==$ "4_2_4_3_4_2_3_4_3_3")
$n.topo.obs[7] <- sum(x.topo.obs ==$ "3_4_2_4_3_3_3_4_2_4")
$n.topo.obs[8] <- sum(x.topo.obs ==$ "4_3_2_4_3_4_2_3_3_4")
$n.topo.obs[9] <- sum(x.topo.obs ==$ "4_4_2_3_2_4_3_4_3_3")
$n.topo.obs[10] <- sum(x.topo.obs ==$ "3_4_4_2_3_3_3_2_4_4")

$n.topo.obs[11] < -sum(x.topo.obs == \text{“}4\_3\_4\_2\_3\_2\_4\_3\_3\_4\text{”})$

$n.topo.obs[12] < -sum(x.topo.obs == \text{“}4\_4\_3\_2\_2\_3\_4\_3\_4\_3\text{”})$

$n.topo.obs[13] < -sum(x.topo.obs == \text{“}3\_3\_3\_3\_2\_4\_4\_4\_4\_2\text{”})$

$n.topo.obs[14] < -sum(x.topo.obs == \text{“}3\_3\_3\_3\_4\_2\_4\_4\_2\_4\text{”})$

$n.topo.obs[15] < -sum(x.topo.obs == \text{“}3\_3\_3\_3\_4\_4\_2\_2\_4\_4\text{”})$

$write(n.topo.obs, file = \text{“}n.topo.obs\text{”}, ncol = 1)$

$n.split.obs < -1 : 10$

$n.split.obs[1] < -n.topo.obs[1] + n.topo.obs[2] + n.topo.obs[3]$

$n.split.obs[2] < -n.topo.obs[4] + n.topo.obs[5] + n.topo.obs[6]$

$n.split.obs[3] < -n.topo.obs[7] + n.topo.obs[8] + n.topo.obs[9]$

$n.split.obs[4] < -n.topo.obs[10] + n.topo.obs[11] + n.topo.obs[12]$

$n.split.obs[5] < -n.topo.obs[1] + n.topo.obs[4] + n.topo.obs[13]$

$n.split.obs[6] < -n.topo.obs[2] + n.topo.obs[7] + n.topo.obs[14]$

$n.split.obs[7] < -n.topo.obs[3] + n.topo.obs[10] + n.topo.obs[15]$

$n.split.obs[8] < -n.topo.obs[5] + n.topo.obs[8] + n.topo.obs[15]$

$n.split.obs[9] < -n.topo.obs[6] + n.topo.obs[11] + n.topo.obs[14]$

$n.split.obs[10] < -n.topo.obs[9] + n.topo.obs[13] + n.topo.obs[12]$

$n.topo.sim < -1 : 15$

$n.split.sim < -1 : 10$

$string < -paste(\text{“}vector\_ts5a.sim\text{”})$

$x.topo.sim < -read.table(string)$

$n.topo.sim[1] < -sum(x.topo.sim == \text{“}2\_3\_4\_4\_3\_4\_4\_3\_3\_2\text{”})$

$n.topo.sim[2] < -sum(x.topo.sim == \text{“}2\_4\_3\_4\_4\_3\_4\_3\_2\_3\text{”})$

$n.topo.sim[3] < -sum(x.topo.sim == \text{“}2\_4\_4\_3\_4\_4\_3\_2\_3\_3\text{”})$

$n.topo.sim[4] < -sum(x.topo.sim == \text{“}3\_2\_4\_4\_3\_3\_3\_4\_4\_2\text{”})$

$n.topo.sim[5] < -sum(x.topo.sim == \text{“}4\_2\_3\_4\_4\_3\_2\_3\_4\_3\text{”})$

$n.topo.sim[6] < -sum(x.topo.sim == \text{“}4\_2\_4\_3\_4\_2\_3\_4\_3\_3\text{”})$

$n.topo.sim[7] < -sum(x.topo.sim == \text{“}3\_4\_2\_4\_3\_3\_3\_4\_2\_4\text{”})$

$n.topo.sim[8] < -sum(x.topo.sim == \text{“}4\_3\_2\_4\_3\_4\_2\_3\_3\_4\text{”})$

$n.topo.sim[9] < -sum(x.topo.sim == \text{“}4\_4\_2\_3\_2\_4\_3\_4\_3\_3\text{”})$

$n.topo.sim[10] < -sum(x.topo.sim == \text{“}3\_4\_4\_2\_3\_3\_3\_2\_4\_4\text{”})$

$n.topo.sim[11] < -sum(x.topo.sim == \text{“}4\_3\_4\_2\_3\_2\_4\_3\_3\_4\text{”})$

$n.topo.sim[12] < -sum(x.topo.sim == \text{“}4\_4\_3\_2\_2\_3\_4\_3\_4\_3\text{”})$

$n.topo.sim[13] < -sum(x.topo.sim == \text{“}3\_3\_3\_3\_2\_4\_4\_4\_4\_2\text{”})$

$n.topo.sim[14] < -sum(x.topo.sim == \text{“}3\_3\_3\_3\_4\_2\_4\_4\_2\_4\text{”})$

$n.topo.sim[15] < -sum(x.topo.sim == \text{“}3\_3\_3\_3\_4\_4\_2\_2\_4\_4\text{”})$

$n.split.sim[1] < -n.topo.sim[1] + n.topo.sim[2] + n.topo.sim[3]$

$n.split.sim[2] < -n.topo.sim[4] + n.topo.sim[5] + n.topo.sim[6]$

$n.split.sim[3]$¡-$n.topo.sim[7] + n.topo.sim[8] + n.topo.sim[9]$

$n.split.sim[4]$¡-$n.topo.sim[10] + n.topo.sim[11] + n.topo.sim[12]$

$n.split.sim[5]$¡-$n.topo.sim[1] + n.topo.sim[4] + n.topo.sim[13]$

$n.split.sim[6]$¡-$n.topo.sim[2] + n.topo.sim[7] + n.topo.sim[14]$

$n.split.sim[7]$¡-$n.topo.sim[3] + n.topo.sim[10] + n.topo.sim[15]$

$n.split.sim[8]$¡-$n.topo.sim[5] + n.topo.sim[8] + n.topo.sim[15]$

$n.split.sim[9]$¡-$n.topo.sim[6] + n.topo.sim[11] + n.topo.sim[14]$

$n.split.sim[10]$¡-$n.topo.sim[9] + n.topo.sim[13] + n.topo.sim[12]$

$D.topo = sum((n.topo.obs - n.topo.sim)^2)$

$D.split = sum((n.split.obs - n.split.sim)^2)$

$write(c(D.topo, D.split), ncol = 2, file = \text{“}Distance.output\text{”}, append = TRUE)$

3. R Code for computing the posterior probability, MLE, and bootstrapping, which is “post.boot.r”.

$library(ape)$

dist< − read.table(“Distance.output”)

st< −read.tree(“st_sim”)

M.topo< −sort(dist$V1)

```
Best.trees.topo< −which(dist$V1<=M.topo[100])
M.split< −sort(dist$V2)
Best.trees.split< −which(dist$V2<=M.split[100])
library(ape)
posterior.topo< −rep(0,7)
posterior.split< −rep(0,7)
for(i in 1:100){


    tree.topo< −st[[Best.trees.topo[i]]]


    treestring.topo< −write.tree(tree.topo)
treestring2.topo< −gsub("//d","",treestring.topo)
treestring3.topo< −gsub(":","",treestring2.topo)
if (treestring3.topo=="((((A.,B.).,C.).,D.).,E.);") posterior.topo[1]
< − posterior.topo[1]+1
if (treestring3.topo=="((((A.,B.).,C.).,E.).,D.);") posterior.topo[2]
< − posterior.topo[2]+1
if (treestring3.topo=="((((D.,E.).,C.).,B.).,A.);") posterior.topo[3]
< − posterior.topo[3]+1
if (treestring3.topo=="((((D.,E.).,C.).,A.).,B.);") posterior.topo[4]
< − posterior.topo[4]+1
if (treestring3.topo=="(((D.,E.).,(A.,B.).).,C.);") posterior.topo[5]
< − posterior.topo[5]+1
if (treestring3.topo=="(((A.,B.).,C.).,(D.,E.).);") posterior.topo[6]
< − posterior.topo[6]+1
if (treestring3.topo=="(((D.,E.).,C.).,(A.,B.).);") posterior.topo[7]
< − posterior.topo[7]+1
```

print(treestring3.topo)

tree.split< −st[[Best.trees.split[i]]]

treestring.split< −write.tree(tree.split)

treestring2.split< −gsub(“//d”,””,treestring.split)

treestring3.split< −gsub(“:”,“”,treestring2.split)

if (treestring3.split==“((((A.,B.).,C.).,D.).,E.);”) posterior.split[1]

< − posterior.split[1]+1

if (treestring3.split==“((((A.,B.).,C.).,E.).,D.);”) posterior.split[2]

< − posterior.split[2]+1

if (treestring3.split==“((((D.,E.).,C.).,B.).,A.);”) posterior.split[3]

< − posterior.split[3]+1

if (treestring3.split==“((((D.,E.).,C.).,A.).,B.);”) posterior.split[4]

< − posterior.split[4]+1

if (treestring3.split==“(((D.,E.).,(A.,B.).).,C.);”) posterior.split[5]

< − posterior.split[5]+1

if (treestring3.split==“(((A.,B.).,C.).,(D.,E.).);”) posterior.split[6]

< − posterior.split[6]+1

if (treestring3.split==“(((D.,E.).,C.).,(A.,B.).);”) posterior.split[7]

< − posterior.split[7]+1

print(treestring3.split) }

write(posterior.topo,file=“posterior.topo”,append=TRUE,ncol=7)

write(posterior.split,file=“posterior.split”,append=TRUE,ncol=7)

Ayed< −read.table(”n.topo.obs”) n.topo.obs< −Ayed$V1

u_1< $-function(tree, x, y, z)\{$

X< $-exp(-x)$

Y< $-exp(-y)$

Z< $-exp(-z)$

value< −0

if(tree==1) value< $-1-(2/3)*X-(2/3)*Y+(1/3)*X*Y+(1/18)*X*Y^3+(1/90)*X*Y^3*Z^6$

if(tree==2) value< $-1-(2/3)*X-(2/3)*Y+(1/3)*X*Y+(1/18)*X*Y^3+(1/90)*X*Y^3*Z^6$

if(tree==3) value< $-1-(2/3)*X-(2/3)*Y+(1/3)*X*Y+(1/18)*X*Y^3+(1/90)*X*Y^3*Z^6$

if(tree==4) value< $-1-(2/3)*X-(2/3)*Y+(1/3)*X*Y+(1/18)*X*Y^3+(1/90)*X*Y^3*Z^6$

if(tree==5) value< $-1-(2/3)*X-(2/3)*Y+(4/9)*X*Y-(2/45)*X*Y*Z^6$

if(tree==6) value< $-1-(2/3)*X-(2/3)*Y*Z+(1/3)*X*Y*Z+(1/15)*X*Y^3*Z$

if(tree==7) value< $-1-(2/3)*X-(2/3)*Y*Z+(1/3)*X*Y*Z+(1/15)*X*Y^3*Z$

return(value)}

u_2< $-function(tree,x,y,z)\{$

X< $-exp(-x)$

Y< $-exp(-y)$

Z< $-exp(-z)$

value< -0

if(tree==1) value< $-(1/3)*Y-(1/6)*X*Y-(1/9)*X*Y^3+(1/90)*X*Y^3*Z^6$

if(tree==2) value< $-(1/3)*Y-(1/6)*X*Y-(1/18)*X*Y^3-(2/45)*X*Y^3*Z^6$

if(tree==3) value< $-(1/3)*X-(1/3)*X*Y+(1/18)*X*Y^3+(1/90)*X*Y^3*Z^6$

if(tree==4) value< $-(1/3)*X-(1/3)*X*Y+(1/18)*X*Y^3+(1/90)*X*Y^3*Z^6$

if(tree==5) value< $-(1/3)*Y-(5/18)*X*Y+(1/90)*X*Y*Z^6$

if(tree==6) value< $-(1/3)*Y*Z-(1/6)*X*Y*Z-(1/10)*X*Y^3*Z$

if(tree==7) value< $-(1/3)*X-(1/3)*X*Y*Z+(1/15)*X*Y^3*Z$

return(value) }

u_3< $-function(tree,x,y,z)\{$

X< $-exp(-x)$

Y< $-exp(-y)$

Z< $-exp(-z)$

value< -0

if(tree==1) value< $-(1/3)*Y-(1/6)*X*Y-(1/18)*X*Y^3-(2/45)*X*Y^3*Z^6$

if(tree==2) value< $-(1/3)*Y-(1/6)*X*Y-(1/9)*X*Y^3+(1/90)*X*Y^3*Z^6$

if(tree==3) value< $-(1/3)*X-(1/3)*X*Y+(1/18)*X*Y^3+(1/90)*X*Y^3*Z^6$

if(tree==4) value< $-(1/3)*X-(1/3)*X*Y+(1/18)*X*Y^3+(1/90)*X*Y^3*Z^6$

if(tree==5) value< $-(1/3)*Y-(5/18)*X*Y+(1/90)*X*Y*Z^6$

if(tree==6) value< $-(1/3)*Y*Z-(1/6)*X*Y*Z-(1/10)*X*Y^3*Z$

if(tree==7) value< $-(1/3)*X-(1/3)*X*Y*Z+(1/15)*X*Y^3*Z$

return(value) }

u_4< $-function(tree,x,y,z)\{$

X< $-exp(-x)$

Y< $-exp(-y)$

Z< $-exp(-z)$

value< -0

if(tree==1) value< $-(1/3)*X-(1/3)*X*Y+(1/18)*X*Y^3+(1/90)*X*Y^3*Z^6$

if(tree==2) value< $-(1/3)*X-(1/3)*X*Y+(1/18)*X*Y^3+(1/90)*X*Y^3*Z^6$

if(tree==3) value< $-(1/3)*Y-(1/6)*X*Y-(1/18)*X*Y^3-(2/45)*X*Y^3*Z^6$

if(tree==4) value< $-(1/3)*Y-(1/6)*X*Y-(1/9)*X*Y^3+(1/90)*X*Y^3*Z^6$

if(tree==5) value< $-(1/3)*X-(5/18)*X*Y+(1/90)*X*Y*Z^6$

if(tree==6) value< $-(1/3)*X-(1/3)*X*Y*Z+(1/15)*X*Y^3*Z$

if(tree==7) value< $-(1/3)*Y*Z-(1/6)*X*Y*Z-(1/10)*X*Y^3*Z$

return(value) }

u_5< $-function(tree,x,y,z)\{$

X< $-exp(-x)$

Y< $-exp(-y)$

Z< $-exp(-z)$

value< -0

if(tree==1) value< $-(1/6) * X * Y - (1/9) * X * Y^3 + (1/90) * X * Y^3 * Z^6$

if(tree==2) value< $-(1/6) * X * Y - (1/18) * X * Y^3 - (2/45) * X * Y^3 * Z^6$

if(tree==3) value< $-(1/18) * X * Y^3 + (1/90) * X * Y^3 * Z^6$

if(tree==4) value< $-(1/18) * X * Y^3 + (1/90) * X * Y^3 * Z^6$

if(tree==5) value< $-(1/18) * X * Y + (1/90) * X * Y * Z^6$

if(tree==6) value< $-(1/6) * X * Y * Z - (1/10) * X * Y^3 * Z$

if(tree==7) value< $-(1/15) * X * Y^3 * Z$

return(value) }

u_6< $-function(tree, x, y, z)\{$

X< $-exp(-x)$

Y< $-exp(-y)$

Z< $-exp(-z)$

value< -0

if(tree==1) value< $-(1/6) * X * Y - (1/18) * X * Y^3 - (2/45) * X * Y^3 * Z^6$

if(tree==2) value< $-(1/6) * X * Y - (1/9) * X * Y^3 + (1/90) * X * Y^3 * Z^6$

if(tree==3) value< $-(1/18) * X * Y^3 + (1/90) * X * Y^3 * Z^6$

if(tree==4) value< $-(1/18) * X * Y^3 + (1/90) * X * Y^3 * Z^6$

if(tree==5) value< $-(1/18) * X * Y + (1/90) * X * Y * Z^6$

if(tree==6) value< $-(1/6) * X * Y * Z - (1/10) * X * Y^3 * Z$

if(tree==7) value< $-(1/15) * X * Y^3 * Z$

return(value) }

u_7< $-function(tree, x, y, z)\{$

$X < -exp(-x)$

$Y < -exp(-y)$

$Z < -exp(-z)$

value< -0

if(tree==1) value< $-(1/18) * X * Y^3 + (1/90) * X * Y^3 * Z^6$

if(tree==2) value< $-(1/18) * X * Y^3 + (1/90) * X * Y^3 * Z^6$

if(tree==3) value< $-(1/6) * X * Y - (1/18) * X * Y^3 - (2/45) * X * Y^3 * Z^6$

if(tree==4) value< $-(1/6) * X * Y - (1/9) * X * Y^3 + (1/90) * X * Y^3 * Z^6$

if(tree==5) value< $-(1/18) * X * Y + (1/90) * X * Y * Z^6$

if(tree==6) value< $-(1/15) * X * Y^3 * Z$

if(tree==7) value< $-(1/6) * X * Y * Z - (1/10) * X * Y^3 * Z$

return(value) }

u_8< $-function(tree, x, y, z)\{$

$X < -exp(-x)$

$Y < -exp(-y)$

$Z < -exp(-z)$

value< -0

if(tree==1) value< $-(1/18) * X * Y^3 + (1/90) * X * Y^3 * Z^6$

if(tree==2) value< $-(1/18) * X * Y^3 + (1/90) * X * Y^3 * Z^6$

if(tree==3) value< $-(1/18) * X * Y^3 + (1/90) * X * Y^3 * Z^6$

if(tree==4) value< $-(1/18) * X * Y^3 + (1/90) * X * Y^3 * Z^6$

if(tree==5) value< $-(1/9) * X * Y - (2/45) * X * Y * Z^6$

if(tree==6) value< $-(1/15) * X * Y^3 * Z$

if(tree==7) value< $-(1/15) * X * Y^3 * Z$

return(value) }

u_9< $-function(tree, x, y, z)\{$

$X < -exp(-x)$

$Y < -exp(-y)$

$Z < -exp(-z)$

value< -0

if(tree==1) value< $-(1/6) * X * Y - (1/18) * X * Y^3 - (2/45) * X * Y^3 * Z^6$

if(tree==2) value< $-(1/6) * X * Y - (1/9) * X * Y^3 + (1/90) * X * Y^3 * Z^6$

if(tree==3) value< $-(1/18) * X * Y^3 + (1/90) * X * Y^3 * Z^6$

if(tree==4) value< $-(1/18) * X * Y^3 + (1/90) * X * Y^3 * Z^6$

if(tree==5) value< $-(1/18) * X * Y + (1/90) * X * Y * Z^6$

if(tree==6) value< $-(1/6) * X * Y * Z - (1/10) * X * Y^3 * Z$

if(tree==7) value< $-(1/15) * X * Y^3 * Z$

return(value) }

u_10< $-function(tree, x, y, z)\{$

$X < -exp(-x)$

$Y < -exp(-y)$

$Z < -exp(-z)$

value< -0

if(tree==1) value< $-(1/18) * X * Y^3 + (1/90) * X * Y^3 * Z^6$

if(tree==2) value< $-(1/18) * X * Y^3 + (1/90) * X * Y^3 * Z^6$

if(tree==3) value< $-(1/6) * X * Y - (1/18) * X * Y^3 - (2/45) * X * Y^3 * Z^6$

if(tree==4) value< $-(1/6) * X * Y - (1/9) * X * Y^3 + (1/90) * X * Y^3 * Z^6$

if(tree==5) value< $-(1/18) * X * Y + (1/90) * X * Y * Z^6$

if(tree==6) value< $-(1/15) * X * Y^3 * Z$

if(tree==7) value< $-(1/6) * X * Y * Z - (1/10) * X * Y^3 * Z$

return(value) }

u_11< $-function(tree, x, y, z)\{$

$X < -exp(-x)$

$Y < -exp(-y)$

$Z < -exp(-z)$

value< -0 if(tree==1) value< $-(1/18) * X * Y^3 + (1/90) * X * Y^3 * Z^6$

if(tree==2) value< $-(1/18) * X * Y^3 + (1/90) * X * Y^3 * Z^6$

if(tree==3) value< $-(1/18) * X * Y^3 + (1/90) * X * Y^3 * Z^6$

if(tree==4) value< $-(1/18) * X * Y^3 + (1/90) * X * Y^3 * Z^6$

if(tree==5) value< $-(1/9) * X * Y - (2/45) * X * Y * Z^6$

if(tree==6) value< $-(1/15) * X * Y^3 * Z$

if(tree==7) value< $-(1/15) * X * Y^3 * Z$

return(value) }

u_12< $-function(tree, x, y, z)\{$

$X < -exp(-x)$

$Y < -exp(-y)$

$Z < -exp(-z)$

value< -0

if(tree==1) value< $-(1/6) * X * Y - (1/9) * X * Y^3 + (1/90) * X * Y^3 * Z^6$

if(tree==2) value< $-(1/6) * X * Y - (1/18) * X * Y^3 - (2/45) * X * Y^3 * Z^6$

if(tree==3) value< $-(1/18) * X * Y^3 + (1/90) * X * Y^3 * Z^6$

if(tree==4) value< $-(1/18) * X * Y^3 + (1/90) * X * Y^3 * Z^6$

if(tree==5) value< $-(1/18) * X * Y + (1/90) * X * Y * Z^6$

if(tree==6) value< $-(1/6) * X * Y * Z - (1/10) * X * Y^3 * Z$

if(tree==7) value< $-(1/15) * X * Y^3 * Z$

return(value) }

u_13< $-function(tree, x, y, z)\{$

$X < -exp(-x)$

$Y < -exp(-y)$

$Z < -exp(-z)$

value< -0

if(tree==1) value< $-(1/3)*X-(1/3)*X*Y+(1/18)*X*Y^3+(1/90)*X*Y^3*Z^6$

if(tree==2) value< $-(1/3)*X-(1/3)*X*Y+(1/18)*X*Y^3+(1/90)*X*Y^3*Z^6$

if(tree==3) value< $-(1/3)*Y-(1/6)*X*Y-(1/9)*X*Y^3+(1/90)*X*Y^3*Z^6$

if(tree==4) value< $-(1/3)*Y-(1/6)*X*Y-(1/18)*X*Y^3-(2/45)*X*Y^3*Z^6$

if(tree==5) value< $-(1/3) * X - (5/18) * X * Y + (1/90) * X * Y * Z^6$

if(tree==6) value< $-(1/3) * X - (1/3) * X * Y * Z + (1/15) * X * Y^3 * Z$

if(tree==7) value< $-(1/3) * Y * Z - (1/6) * X * Y * Z - (1/10) * X * Y^3 * Z$

return(value) }

u_14< $-function(tree, x, y, z)\{$

$X < -exp(-x)$

$Y < -exp(-y)$

$Z < -exp(-z)$

value< -0

if(tree==1) value$< -(1/18) * X * Y^3 + (1/90) * X * Y^3 * Z^6$

if(tree==2) value$< -(1/18) * X * Y^3 + (1/90) * X * Y^3 * Z^6$

if(tree==3) value$< -(1/6) * X * Y - (1/9) * X * Y^3 + (1/90) * X * Y^3 * Z^6$

if(tree==4) value$< -(1/6) * X * Y - (1/18) * X * Y^3 - (2/45) * X * Y^3 * Z^6$

if(tree==5) value$< -(1/18) * X * Y + (1/90) * X * Y * Z^6$

if(tree==6) value$< -(1/15) * X * Y^3 * Z$

if(tree==7) value$< -(1/6) * X * Y * Z - (1/10) * X * Y^3 * Z$

return(value) }

u_15$< -function(tree, x, y, z)\{$

$X < -exp(-x)$

$Y < -exp(-y)$

$Z < -exp(-z)$

value< -0

if(tree==1) value$< -(1/18) * X * Y^3 + (1/90) * X * Y^3 * Z^6$

if(tree==2) value$< -(1/18) * X * Y^3 + (1/90) * X * Y^3 * Z^6$

if(tree==3) value$< -(1/6) * X * Y - (1/9) * X * Y^3 + (1/90) * X * Y^3 * Z^6$

if(tree==4) value$< -(1/6) * X * Y - (1/18) * X * Y^3 - (2/45) * X * Y^3 * Z^6$

if(tree==5) value$< -(1/18) * X * Y + (1/90) * X * Y * Z^6$

if(tree==6) value$< -(1/15) * X * Y^3 * Z$

if(tree==7) value$< -(1/6) * X * Y * Z - (1/10) * X * Y^3 * Z$

return(value) }

log.Like$< -function(tree, x, y, z, n.topo.obs)\{$

value$< -$ n.topo.obs[1]*log(u_1(tree,x,y,z))+ n.topo.obs[2]*log(u_2(tree,x,y,z))+

n.topo.obs[3]*log(u_3(tree,x,y,z))+ n.topo.obs[4]*log(u_4(tree,x,y,z))+

```
n.topo.obs[5]*log(u_5(tree,x,y,z))+ n.topo.obs[6]*log(u_6(tree,x,y,z)) +
n.topo.obs[7]*log(u_7(tree,x,y,z))+ n.topo.obs[8]*log(u_8(tree,x,y,z))+
n.topo.obs[9]*log(u_9(tree,x,y,z)) + n.topo.obs[10]*log(u_10(tree,x,y,z))+
n.topo.obs[11]*log(u_11(tree,x,y,z))+ n.topo.obs[12]*log(u_12(tree,x,y,z)) +
n.topo.obs[13]*log(u_13(tree,x,y,z))+ n.topo.obs[14]*log(u_14(tree,x,y,z))+
n.topo.obs[15]*log(u_15(tree,x,y,z))
return(value) }
best.MLE< −c(0,0,0,0,0)
x < −-log(seq(0.01,1,0.01))
y < −x
z < −x
best.MLE[4]< −log.Like(1, 1, 1, 1, n.topo.obs)
for (h in 1:7){
for(i in 1:length(x)){
print(c(h,i))
for(j in 1:length(y)){
for(k in 1:length(z)){
temp< −log.Like(h,x[i],y[j],z[k],n.topo.obs)
if(temp>best.MLE[4]){
best.MLE[4]< −temp
best.MLE[1]< −i
best.MLE[2]< −j
best.MLE[3]< −k
best.MLE[5]< −h
} } } } }


    boot.trees< −NULL
for(i in 1:15) {
```

```
boot.trees <- c(boot.trees,rep(i,n.topo.obs[i])) }
best<-c(0,0,0,0,0)
x<- -log(seq(0.01,1,0.01))
y <- -x
z <- -x
boot.suport<- -rep(0,7)
for (b in 1:50){
boot.trees.temp<- - sample(boot.trees,replace=TRUE)
n.topo.boot<- -1:15
n.topo.boot[1]<- -sum(boot.trees.temp==1)
n.topo.boot[2]<- -sum(boot.trees.temp==2)
n.topo.boot[3]<- -sum(boot.trees.temp==3)
n.topo.boot[4]<- -sum(boot.trees.temp==4)
n.topo.boot[5]<- -sum(boot.trees.temp==5)
n.topo.boot[6]<- -sum(boot.trees.temp==6)
n.topo.boot[7]<- -sum(boot.trees.temp==7)
n.topo.boot[8]<- -sum(boot.trees.temp==8)
n.topo.boot[9]<- -sum(boot.trees.temp==9)
n.topo.boot[10]<- -sum(boot.trees.temp==10)
n.topo.boot[11]<- -sum(boot.trees.temp==11)
n.topo.boot[12]<- -sum(boot.trees.temp==12)
n.topo.boot[13]<- -sum(boot.trees.temp==13)
n.topo.boot[14]<- -sum(boot.trees.temp==14)
n.topo.boot[15]<- -sum(boot.trees.temp==15)
best[4]<- -log.Like(1,1,1,1,n.topo.boot)
for (h in 1:7){
for(i in 1:length(x)){
print(c(b,h,i))
```

for(j in 1:length(y)){
for(k in 1:length(z)){
temp.boot< −log.Like(h,x[i],y[j],z[k],n.topo.boot)
if(temp.boot>best[4]){
best[4]¡-temp.boot
best[1]< $-i$
best[2]< $-j$
best[3]< $-k$
best[5]< $-h$
} } } } }
boot.suport[best[5]]< −boot.suport[best[5]]+1
print(best)}
write(c(best.MLE,boot.suport),ncol=12,file=“output.MLE.Bootstrap.r.1”)

A.2

Script to simulate the DNA sequence and calculate the ABC method, MLE, and Bootstrapping, which is scriptABC.
cd name direction
$for((h = 1; h$¡= $1; h++))$
do
rm -f vector_ts5a.obs
rm -f vector_ts5a.sim
rm -f st_sim
rm -f Distance.output
./hybrid-Lambda -spcu ABCDE1 -num 100 -seed 217
sed ’s/_1//g’ OUT_coal_unit > data-ts5a

$for((i = 1; i <= 100; i++))$
do
head -$i data-ts5a | tail -1 > gt
./seq-gen -l1000 -s.005 -mGTR -a1.0 -g4 -i.1 -f.3,.2,.2,.3 -z19$h5$i -op < gt > foo #
This command is used to generate the DNA sequences from the observed data.
./phyml -ifoo -mGTR -fe -ve -c4 -ae –r_seed 5ih –no_memory_check –quiet > log #
Also this command is used to compute the DNA sequences from the observed data.
#./phyml foo 0 s 1 0 GTR e e 1 1.0 BIONJ y y
R CMD BATCH root.r # This R code uses to rooted gene trees from DNA sequences.
./vector gt.2 0 >> vector_ts5a.obs # This command is used to run the vector program to compute the topology from DNA sequence observed data.
done
$for((j = 1; j <= 50000; j++))$
do
rm -f vector_ts5a.sim
R CMD BATCH priorst.r
./ultrametric tempst.1 > ts5a.new
cat ts5a.new >> st_sim
echo $h $j >> log-output-number
./hybrid-Lambda -spcu ts5a.new -num 100 -seed 2171
sed 's/_1//g' OUT_coal_unit > data-ts5a.new
$for((k = 1; k <= 100; k++))$
do
head -$k data-ts5a.new | tail -1 > gt.1
./seq-gen -l1000 -s.005 -mGTR -a1.0 -g4 -i.1 -f.3,.2,.2,.3 -z19$h05$j5$k -op < $gt.1$ >
foo # This command is used to generate DNA sequences from the simulated data.
#./phyml -ifoo -mGTR -fe -ve -c4 -ae –r_seed 5kh337$j –no_memory_check –quiet
> log Also this comment to compute DNA sequences from the simulated data.

./phyml foo 0 s 1 0 GTR e e 1 1.0 BIONJ y y
R CMD BATCH root.r # This R code uses to rooted gene trees from DNA sequences.
./vector gt.2 0 >> vector_ts5a.sim # This command is used to run the vector program to compute the topology from DNA sequence simulated data.
done
R CMD BATCH compare.r.2
done
R CMD BATCH post.boot.r
done

# This R code uses to rooted the gene tree that gets from the DNA sequence simulation, which is " root.r"
library(ape)
a< −read.tree("foo_phyml_tree.txt") # This code read the output from Phyml program
b< −multi2di(a) # This to rooted the gene trees and create the gt.2 file.
write.tree(b,"gt.2")

A.3

# R code to compute the distance between the simulated data and observed data for the eight taxa split, which is "compare.r.3.1"
library("sets")
x.topo.obs < − read.table("splits_ts8s.obs3.1")
n.obs.topo< − gset(x.topo.obs)
names(n.obs.topo)
n.obs.topo< − table(x.topo.obs)

topo.obs.n< − matrix(n.obs.topo)
n.topo.obs< −gset(sort(unique(x.topo.obs$V1)),memberships=topo.obs.n[,1])
x.topo.sim < − read.table(“splits_ts8s.sim3.1”)
n.sim.topo< − gset(x.topo.sim)
names(n.sim.topo)
n.sim.topo< − table(x.topo.sim)
topo.sim.n< − matrix(n.sim.topo)
n.topo.sim< −gset(sort(unique(x.topo.sim$V1)),memberships=topo.sim.n[,1])
topo.D1= n.topo.obs-n.topo.sim | n.topo.sim-n.topo.obs
topo.D2< − gset_memberships(topo.D1)
D.topo=sum(topo.D2)
write (c(D.topo), ncol=1, file=“Output.8.taxa3.1”, append=TRUE)

# R code to find the best 100 trees for using to compute the posterior probability.
library(ape)
dist.3.1< − read.table(“Output.8.taxa3.1”)
st.3.1< −read.tree(“st_sim3.1”)
M.topo3.1< −sort(dist.3.1$V1)
Best.trees.topo3< −which(dist.3.1$V1<=M.topo3.1[100])
$for(iin1:100)${
tree.topo3< −st.3.1[[Best.trees.topo3[i]]]
write.tree(tree.topo3,“split.8.taxa3.1”, append=TRUE)}

Script for 8 taxa.
Make a temp directory on /dev/shm where small temp files will be written and read
from TEMP=/dev/shm/$1
mkdir $TEMP

Go to directory and clean up old files

cd $HOME/$2 || exit

rm -f split.8.taxa3.1

rm -f Output.8.taxa3.1

$for((h = 1; h <= 1; h++))$

do

rm -f splits_ts8s.obs3.1

rm -f splits_ts8s.sim3.1

rm -f st_sim3.1

Run the simulation

hybrid-Lambda -spcu ST8cat.txt -num 100 -seed $1 # This command is used run hybrid-Lambda to simulate the observed data for the 8-taxa.

sed 's/_1//g' OUT_coal_unit > data-ts8s3.1

$for((i = 1; i <= 100; i++))$

do

head -$i data-ts8s3.1 | tail -1 > $TEMP/gt

splits $TEMP/gt 0 > $TEMP/temp83.1 # This command is used to run the split program to get the split of the 8-taxa of the observed.

head -5 $TEMP/temp83.1 >> splits_ts8s.obs3.1

done

$for((j = 1; j <= 12500; j++))$

do

rm -f splits_ts8s.sim3.1

R CMD BATCH priorst.r.3.1 # This command is used to run the R code to create data from the prior.

ultrametric tempst.3.1 > $TEMP/ts8s.new3.1 # make the tree that got from prior clock's.

cat $TEMP/ts8s.new3.1 >> st_sim3.1

echo $h $j >> log-output-number

hybrid-Lambda -spcu $TEMP/ts8s.new3.1 -num 100 -seed 1j # This command is used to run hybrid-Lambda to get the simulated data for the 8-taxa from the prior.

sed 's/_1//g' OUT_coal_unit > $TEMP/data-ts8s.new3.1

$for((k = 1; k <= 100; k++))$

do

head -$k $TEMP/data-ts8s.new3.1 | tail -1 > $TEMP/gt.1

splits $TEMP/gt.1 0 > $TEMP/temp8.3.1 # This command is used to run the split program to get the split of the 8-taxa of the simulated.

head -5 $TEMP/temp8.3.1 >> splits_ts8s.sim3.1

done

R CMD BATCH compare.r.3.1 # This command is used to run the R code to compute the distance between the split of observed data and split of simulated data.

done

R CMD BATCH post.boot.r.3.1 # This command is used to run the R code to compute the smallest distance of 100 trees.

done

Clean up the temp directory on /dev/shm

mkdir temp_files

cp $TEMP/* ./temp_files

rm -r $TEMP

cp split.8.taxa3.1 intree # cp means cope the file "split.8.taxa3.1", a smallest distance, which got from this code file "post.boot.r.3.1" , to file intree to use as input data for consensus program to compute the posterior probability.

./consense <<EOF # The command is used to run consensus program. The EOF command is used to open the file and apply the following this command.

R # This command is used to ask the consensus program to choose the rooted the

output.
Y # means Yes.
EOF # end the file.
cat outfile >> outfile1 # To save the shape that gets from the consensus program in this file "outfile1".
cat outtree >>outtree1 # To save the output from the consensus program in this file "outtile1".
rm -f outfile
cat ts8bal outtree > intree # cope the original data to outtree file to use as input data for treedist program.
./treedist << EOF # This command is used to run the treedist program to compute the distance between the true tree and the simulated tree.
R # Ask the treedist program to root the tree.
D # To give the distance between the trees.
Y # Yes
EOF # end the command.
cat outfile >> outfile8 # Save the output from the treedist in this file "outfile8".

A.4

Code for computing the MLE for rooted vs unrooted trees with DNA and without DNA.

The code is from data.create.r
library(*TreeSim*) # This code is used for running the library TreeSim to use the function to create species trees.
x< −sim.bd.taxa(n-taxa,number of species,λ,$\lambda\times\mu$, complete=FALSE) # This code

is used for running the sim.bd.taxa function to simulate the species tree, but it is necessary to specify the number of taxa desired, the number of gene trees desired, the value of λ desired, and the last entry of the value of μ times the λ.

$for(iin1:50)\{$

filename< −paste("speciestree.8-0.1-0") # To save the data generated after adding the outgroup.

filename1< −paste("speciestree.without.outgroup") # To save the data without out-group.

string< −write.tree(x[[i]])

g< −gregexpr(")", string, fixed=TRUE)

loc< −g[[1]]

string1< −substring(string,1,loc[length(loc)])

string2< −paste("(",string1,":10,z:10);", sep="") # The code is from string until string2 used for adding the outgroup to the species tree.

string3< −paste(string1,";", sep="")

write(string2, file=filename, append=TRUE, ncol=1)

write(string3, file=filename1, append=TRUE, ncol=1)}

The code is from the drop.out-group.

$library(ape)$ # To run the APE library (Analyses of Phylogenetics and Evolution)

Tr¡-read.tree("data-ts8s3.1") # To read the file from the hybrid-Lambda program

$for(iin1:length(Tr))\{$

y< −drop.tip(Tr[[i]], "z") # This code is to drop the outgrop from the species after simulating the gene tree without DNA.

write.tree(y, file="species.unroot", append=TRUE)}

The code is from the drop.out-group.1

library(*ape*)
Tre< −read.tree(“foo_phyml_tree.txt”) # Read the species tree after the gene tree with DNA.
g< −drop.tip(Tre, “z”) ”) # This code is for dropping the outgroup from the species after simulating the gene tree with DNA.
write.tree(g, file=“gt.drop”)

The code is from root.r
library(*ape*)
l*ibrary*(*phytools*) # Run Phytools library to the rooted species on the place desired.
a< −read.tree(“foo_phyml_tree.txt”)
index< −which(a$tip.label==“z”)
b< −reroot(a,index,0.1) # Read the species tree after simulating the DNA until it is rerooted to make the outgroup “z”.
write.tree(b,“gt.2”)

#The code is from PhyML-head.
#NEXUS
BEGIN TREES;

The code is from PhyML.r
x< −read.table(“species.unroot”) # read the species tree without the outgroup after simulating the known gene tree.
$for(i in 1 : 100)\{$
string1< − paste(“gt”,i,sep=“”)
string < − paste(“Tree”,string1,“=”,“[&U]”,x$V1[i])
write(string, file=“phyml-middle”, append=TRUE)

} # The code inside the loop creates the formula of the unrooted gene tree in order to use PhyloNet.
The code is from PhyML.r.1
x< −read.table("data-ts8s3.1") ") # Read the species tree with the outgroup after simulating the known gene tree.
$for(iin1:100)\{$ string1 < − paste("gt",i,sep="")
string < − paste("Tree",string1,"=",x$V1[i])
write(string, file="phyml-middle.1", append=TRUE)
} # The code inside the loop creates the formula of the unrooted gene tree in order to use PhyloNet.

#The code is from PhyML-tail
END;
BEGIN PHYLONET;
InferNetwork_ML (all) 0 -x 5 -m 1000;
END;
# This PhyML-tail code is used for researching the MLE tree by using x=5 (which how many times the search was run), and m=1000 which is the highest value of examinable network topologies.

All code from the PhyML-head until the PhyML-tail makes the format run the PhyloNet program.

Below is the script used for computing the MLE for rooted vs unrooted trees with DNA and without DNA.
rm -f MLE.8.taxa.*rooted* # This command is for removing those files.
R CMD BATCH data.create.r # This command is for running the R code to create

the species trees.

$for((h = 1; h <= 50; h++))$

do

rm -f data-ts8s3.1

rm -f species.unroot

rm -f phyml-middle

rm -f temp.net

rm -f spe.drop.outgp

rm -f data.rooted # All those files need to be removed before starting.

head -$h speciestree.8-0.1-0 | tail -1 > TREE1 #the command to head only the last line from the species.

./ultrametric TREE1 > TREEMLE # This command is for making the species tree ultra-metric.

./hybrid-Lambda -spcu TREEMLE -num 100 # This command is for running the hybrid-lambda program to simulate know gene trees.

sed 's/_1//g' OUT_coal_unit > data-ts8s3.1

R CMD BATCH drop.out-group # This command is for running the R code to drop the outgroup.

R CMD BATCH phyml.r # This command is for running some R codes to create the PhyML-middle unrooted gene tree.

cat phyml-head phyml-middle phyml-tail > temp.net # cope all those file to one file to be input for PhyloNet program.

java -jar PhyloNet_3.6.0.jar temp.net > taxa.8.MLE.output.unrooted # This command is for running the PhyloNet program.

tail -2 taxa.8.MLE.output.unrooted | head -1 >> MLE.8.taxa.unrooted # save the last line to file with append the result.

rm -f phyml-middle.1

R CMD BATCH phyml.r.1 # This command is for running some R codes to create

PhyML-middle.1.rooted gene tree.

cat phyml-head phyml-middle.1 phyml-tail > temp.net.1

java -jar PhyloNet_3.6.0.jar temp.net.1 > taxa.8.MLE.output.rooted # This command is for running the PhyloNet program.

tail -2 taxa.8.MLE.output.rooted | head -1 >> MLE.8.taxa.rooted

$for((i = 1; i <= 100; i++))$

do

head -$i data-ts8s3.1 | tail -1 > gt

./seq-gen -l1000 -s.005 -mGTR -a1.0 -g4 -i.1 -f.3,.2,.2,.3 -op < gt > foo # This command is for running the seq-gen program to create DNA sequences.

./phyml foo 0 s 1 0 GTR e e 1 1.0 BIONJ y y # This command is for running the PhyML program, which is the result of seq-gen used as the input for PhyML to create the estimate gene tree.

R CMD BATCH drop.out-group.1 group # This command is for running the R code to drop the outgroup.

R CMD BATCH root.r #This command is for making the species from DNA rooted.

cat gt.drop >> spe.drop.outgp

cat gt.2 >> data.rooted

done

rm -f phyml-middle.2

R CMD BATCH phyml.r.2 # This command is for running some R codes to create PhyML-middle.2 rooted gene tree.

cat phyml-head phyml-middle.2 phyml-tail > temp.net.2

java -jar PhyloNet_3.6.0.jar temp.net.2 > taxa.8.MLE.output.rootedDNA # This command is for running the PhyloNet program.

tail -2 taxa.8.MLE.output.rootedDNA | head -1 >> MLE.8.taxa.rootedDNA

rm -f phyml-middle.3

R CMD BATCH phyml.r.3 # This command is for running some R codes to create

PhyML-middle unrooted gene tree.
cat phyml-head phyml-middle.3 phyml-tail > temp.net.3
java -jar PhyloNet_3.6.0.jar temp.net.3 > taxa.8.MLE.output.unrootedDNA # This command is for running the PhyloNet program.
tail -2 taxa.8.MLE.output.unrootedDNA — head -1 >> MLE.8.taxa.unrootedDNA
done
$for((i = 1; i <= 50; i++))$
do
head -$i MLE.8.taxa.unrooted | tail -1 > Temp1
head -$i speciestree.without.outgroup | tail -1 > Temp2
These two commands above are for reading the specific line from "MLE.8.taxa.unrooted", which is the result from the PhyloNet program, to compare it with "speciestree.without.outgroup", based on the loop and then save it to these two files "Temp1" and "Temp2"
rm -f outfile
rm -f intree # every time it is necessary to remove the two files above.
cat Temp1 > intree
cat Temp2 >>intree # copy the two files "Temp1" and "Temp2" to be input for the treedist program.
./treedist <<EOF # This command is for running the treedist program.
D # mean distance.
R # rooted.
Y # Yes.
EOF
grep and outfile >> output.unroot.dist # After running the treedist program, save the result for unrooted gene trees without DNA to this file "output.unroot.dist" and also do this with URGT with DNA and RGT with and without DNA.
done

Appendix B

Output for ABC method

B.1

Table B.1: Species 1; ((((A:1,B:1.0):0.1,C:1.1):0.1,D:1.2):0.1,E:1.3);

Number	Posterior Probability of Topology Trees							Posterior Probability of Split Trees						
	1	2	3	4	5	6	7	1	2	3	4	5	6	7
1	70	14	0	4	6	2	4	71	4	0	11	5	4	5
2	31	17	11	8	12	18	3	15	19	18	19	16	12	1
3	39	54	0	0	4	2	1	41	34	10	2	11	2	0
4	75	22	0	0	0	2	1	64	28	0	0	2	5	1
5	11	19	0	1	21	11	37	34	28	4	0	9	13	12
6	93	6	0	0	0	1	0	80	20	0	0	0	0	0
7	60	26	2	2	5	3	2	33	37	10	2	16	2	0
8	51	23	0	0	14	0	12	40	26	1	1	20	0	12
9	41	7	0	6	20	22	4	39	14	1	14	24	5	3
10	46	35	0	0	2	15	2	47	36	0	0	1	12	4
11	66	17	1	1	0	15	0	72	16	3	1	0	8	0
12	34	4	0	0	32	24	6	36	7	0	2	30	18	7
13	58	26	0	0	2	14	0	67	24	0	1	6	2	0
14	74	21	0	0	0	5	0	62	19	0	0	10	9	0
15	57	42	0	0	1	0	0	46	34	1	1	17	0	1
16	10	10	31	36	5	0	8	9	20	26	27	3	0	15
17	24	76	0	0	0	0	0	39	58	0	0	2	1	0
18	8	45	12	0	24	11	0	7	41	21	0	25	6	0
19	82	16	1	0	0	1	0	55	33	1	3	5	2	1
20	88	10	0	0	0	1	1	68	25	2	1	1	1	2
21	36	18	11	6	14	12	3	32	24	18	6	14	5	1
22	45	47	3	1	3	1	0	50	45	4	0	1	0	0
23	92	8	0	0	0	0	0	90	7	0	0	3	0	0
24	31	48	0	0	2	17	2	36	33	0	2	4	21	4
25	62	27	1	0	6	4	0	56	14	2	3	17	8	0
26	32	27	0	5	24	7	5	29	39	1	7	14	9	1
27	54	19	0	1	4	14	8	46	4	3	8	9	16	14
28	63	25	0	1	0	11	0	52	28	2	0	7	11	0
29	79	18	0	0	0	3	0	49	40	1	0	6	4	0
30	26	52	0	0	9	8	5	31	42	1	1	12	8	5
31	87	9	0	0	2	2	0	67	14	4	2	9	2	2
32	37	7	0	0	28	20	8	29	14	1	0	38	14	4
33	43	52	0	0	1	0	4	44	52	0	1	1	2	0
34	92	6	0	0	0	2	0	79	19	0	0	0	2	0
35	12	82	0	0	2	4	0	25	65	0	0	2	7	1
36	52	34	0	0	0	14	0	40	20	7	10	10	13	0
37	39	58	0	0	2	0	1	47	38	0	0	15	0	0
38	95	5	0	0	0	0	0	94	4	1	0	0	1	0
39	51	9	3	17	12	8	0	25	20	4	17	21	13	0
40	71	9	2	3	7	7	1	60	14	6	3	7	7	3
41	7	1	14	18	53	4	3	18	1	13	17	41	4	6
42	30	12	19	15	8	10	6	9	5	36	28	18	0	4
43	59	36	0	0	0	5	0	63	31	2	1	0	3	0
44	79	6	4	3	0	5	3	64	9	8	7	5	2	5
45	95	4	0	0	0	1	0	83	10	1	0	4	2	0
46	45	16	3	1	12	19	4	55	23	4	1	6	11	0
47	72	11	0	0	3	13	1	53	25	2	1	6	12	1
48	79	17	0	0	1	3	0	68	20	0	0	7	1	4
49	57	13	2	1	14	13	0	38	20	4	2	21	13	2
50	41	32	0	0	8	19	0	41	34	0	0	13	12	0

Table B.3: Species 2; ((((A:1,B:1.0):0.1,C:1.1):0.1,D:1.2):1.0,E:2.2);

Number	Posterior Probability of Topology Trees							Posterior Probability of Split Trees						
	1	2	3	4	5	6	7	1	2	3	4	5	6	7
1	79	19	0	0	1	1	0	73	26	1	0	0	0	0
2	79	19	0	0	0	2	0	67	30	0	0	0	3	0
3	77	4	0	0	5	14	0	52	10	4	5	16	8	5
4	77	21	0	0	0	2	0	64	31	0	0	5	0	0
5	93	7	0	0	0	0	0	84	16	0	0	0	0	0
6	81	14	0	0	2	3	0	61	28	0	0	8	3	0
7	81	17	0	0	0	2	0	75	24	0	0	0	1	0
8	60	6	1	1	11	20	1	69	7	1	5	8	10	0
9	80	18	0	0	0	2	0	55	33	0	3	7	2	0
10	69	31	0	0	0	0	0	61	37	0	0	1	1	0
11	82	14	0	0	1	1	2	76	20	1	0	0	2	1
12	89	11	0	0	0	0	0	54	42	0	0	1	1	2
13	56	38	0	0	0	6	0	58	37	0	0	1	4	0
14	55	13	4	8	4	15	1	71	14	6	5	2	2	0
15	80	14	0	0	0	6	0	69	25	0	0	0	6	0
16	62	30	0	0	6	2	0	58	30	0	0	8	3	1
17	81	14	0	0	3	2	0	64	13	9	1	11	1	1
18	66	24	0	0	9	1	0	62	26	0	0	12	0	0
19	78	15	0	0	7	0	0	76	15	0	0	5	4	0
20	73	22	0	0	0	5	0	56	32	1	2	2	5	2
21	76	15	0	0	1	8	0	49	33	0	0	9	9	0
22	87	13	0	0	0	0	0	73	27	0	0	0	0	0
23	82	12	0	0	0	6	0	72	20	0	0	1	7	0
24	71	29	0	0	0	0	0	55	43	0	0	1	0	1
25	60	13	2	9	8	8	0	43	21	4	11	6	12	3
26	72	9	4	4	8	1	2	75	19	2	2	2	0	0
27	18	27	0	45	3	4	3	23	29	2	33	7	3	3
28	61	26	0	0	5	5	3	62	20	0	0	16	2	0
29	55	31	0	2	11	1	0	49	41	0	4	6	0	0
30	56	40	0	0	0	4	0	60	29	1	0	7	3	0
31	91	7	0	0	0	2	0	76	21	0	0	0	3	0
32	57	8	0	0	7	28	0	55	20	0	2	5	18	0
33	72	17	0	0	9	0	2	61	26	0	3	9	0	1
34	84	14	0	0	1	0	1	83	14	0	0	3	0	0
35	95	2	0	0	1	2	0	80	11	0	0	2	7	0
36	79	17	0	0	0	4	0	66	31	0	0	0	3	0
37	25	33	3	4	21	10	4	28	24	8	10	19	6	5
38	51	44	0	0	4	1	0	50	36	0	0	13	1	0
39	68	23	0	0	7	2	0	55	33	2	0	10	0	0
40	49	28	0	3	6	7	7	50	34	0	4	3	5	4
41	61	33	0	1	1	4	0	65	33	0	0	1	1	0
42	65	32	0	0	0	3	0	65	24	0	0	0	11	0
43	78	16	0	0	0	6	0	74	17	1	2	0	6	0
44	93	6	0	0	0	1	0	80	14	0	0	2	4	0
45	40	14	0	1	35	8	2	34	19	0	4	29	5	9
46	48	51	0	0	0	1	0	51	48	0	0	1	0	0
47	58	37	0	0	0	5	0	49	46	0	0	1	4	0
48	30	35	0	3	15	14	3	19	61	1	3	12	3	1
49	50	20	0	1	7	12	10	39	32	4	1	4	11	9
50	53	37	1	0	3	4	2	34	60	0	0	4	2	0

Table B.5: Species 3; ((((A:1,B:1.0):1.0,C:2.0):0.1,D:2.1):0.1,E:2.2);

Number	Posterior Probability of Topology Trees							Posterior Probability of Split Trees						
	1	2	3	4	5	6	7	1	2	3	4	5	6	7
1	59	27	0	0	0	4	10	58	25	0	1	5	3	8
2	34	38	2	2	14	1	9	40	32	0	0	18	3	7
3	47	21	2	3	6	9	12	42	42	0	0	4	7	5
4	1	94	0	0	4	0	1	4	87	0	0	7	1	1
5	29	5	0	0	56	2	8	38	12	0	0	45	1	4
6	39	50	0	0	7	4	0	33	54	0	0	9	3	1
7	39	48	0	0	6	1	6	31	56	1	0	8	0	4
8	38	21	2	4	9	25	1	42	25	3	1	9	20	0
9	26	36	5	2	20	6	5	23	31	2	4	35	2	3
10	47	35	0	0	14	1	3	37	34	0	0	24	1	4
11	58	32	0	0	6	3	1	59	29	0	0	6	4	2
12	65	24	0	0	7	4	0	45	26	2	0	18	1	8
13	33	42	0	0	7	7	11	20	38	1	2	13	10	16
14	3	53	6	5	7	3	23	4	54	4	3	8	6	21
15	57	24	0	0	19	0	0	49	26	0	0	25	0	0
16	29	39	1	3	21	6	1	39	31	1	0	26	0	3
17	44	36	0	0	16	3	1	49	20	0	0	23	3	5
18	31	60	0	0	3	0	6	27	45	0	0	8	3	17
19	56	22	0	0	11	11	0	54	9	0	0	15	13	9
20	48	34	0	0	7	3	8	45	41	0	0	7	1	6
21	35	20	0	0	34	6	5	31	29	0	0	32	5	3
22	49	35	0	0	11	1	4	39	33	0	0	27	1	0
23	41	9	0	1	37	6	6	29	13	1	1	39	4	13
24	21	25	0	0	37	10	7	32	32	0	0	16	9	11
25	96	1	0	0	3	0	0	90	2	0	0	4	4	0
26	37	27	1	1	21	5	8	28	24	2	4	26	2	14
27	30	41	0	0	18	4	7	26	40	0	0	18	2	14
28	35	50	0	0	5	3	7	43	49	0	0	3	3	2
29	47	12	0	0	39	0	2	43	26	0	0	28	2	1
30	64	32	0	0	2	0	2	57	28	0	0	13	2	0
31	75	17	0	0	2	0	6	72	22	0	0	4	0	2
32	36	49	0	0	13	0	2	35	47	0	0	13	0	5
33	24	41	0	0	17	2	16	40	34	0	0	18	1	7
34	37	30	9	6	11	0	7	33	22	16	6	14	0	9
35	87	0	0	0	8	5	0	65	0	0	0	34	1	0
36	38	26	0	0	26	7	3	23	40	0	0	27	5	5
37	41	22	0	0	23	4	10	28	35	0	0	22	4	11
38	32	25	0	0	40	0	3	19	12	0	0	61	1	7
39	39	48	0	0	8	0	5	54	35	0	0	10	0	1
40	17	49	0	0	13	10	11	15	38	0	1	16	17	13
41	41	43	0	0	14	0	2	27	24	0	1	33	2	13
42	19	46	0	0	24	6	5	27	43	0	0	24	4	2
43	33	39	0	1	14	12	1	44	27	0	0	18	8	3
44	46	38	0	0	10	2	4	47	29	0	0	13	5	6
45	11	80	0	0	3	4	2	9	60	0	0	26	4	1
46	33	36	0	0	15	3	13	19	38	0	0	19	3	21
47	57	24	0	0	9	2	8	48	27	0	0	13	3	9
48	23	47	0	0	29	1	0	32	39	0	0	29	0	0
49	26	31	0	0	30	0	13	36	38	0	0	21	0	5
50	52	16	0	0	18	11	3	55	13	0	0	20	10	2

Table B.7: Species 4; ((((A:1,B:1.0):0.1,C:1.1):1.0,D:2.1):0.1,E:2.2);

Number	Posterior Probability of Topology Trees							Posterior Probability of Split Trees						
	1	2	3	4	5	6	7	1	2	3	4	5	6	7
1	30	26	10	11	9	13	1	35	33	17	4	7	4	0
2	22	28	0	0	0	50	0	20	33	0	0	0	47	0
3	29	34	0	0	0	37	0	38	35	0	0	0	27	0
4	25	48	0	0	3	24	0	26	42	0	3	5	24	0
5	30	42	0	0	0	28	0	26	28	0	0	10	35	1
6	31	21	0	0	0	48	0	27	22	3	2	0	46	0
7	39	37	0	0	0	24	0	45	39	0	0	0	16	0
8	33	35	0	0	0	32	0	33	37	0	1	0	29	0
9	25	39	0	0	0	36	0	32	39	0	0	0	29	0
10	44	34	0	0	0	22	0	35	44	0	0	0	21	0
11	42	37	0	0	0	21	0	40	30	0	0	0	30	0
12	46	42	0	0	2	10	0	38	39	0	1	5	17	0
13	34	31	0	0	2	33	0	49	26	0	0	5	20	0
14	49	16	0	0	2	33	0	46	15	5	0	11	19	4
15	39	33	0	0	0	28	0	29	30	0	0	0	41	0
16	52	30	0	0	0	18	0	49	32	1	0	0	18	0
17	46	32	0	0	1	21	0	30	27	0	1	6	36	0
18	24	17	1	1	5	52	0	26	34	1	0	2	37	0
19	21	18	0	0	1	60	0	13	12	0	2	3	69	1
20	46	34	0	0	0	20	0	47	31	0	0	0	22	0
21	44	20	0	0	0	36	0	47	16	0	0	0	37	0
22	24	37	0	0	0	39	0	29	39	2	0	1	29	0
23	25	34	0	0	17	23	1	26	26	1	0	17	29	1
24	34	20	0	0	0	46	0	29	32	0	0	0	39	0
25	23	33	0	0	5	39	0	29	36	2	0	5	28	0
26	38	29	0	0	0	33	0	38	29	0	0	0	33	0
27	46	41	0	0	0	13	0	52	34	0	0	1	12	1
28	27	18	0	0	0	55	0	30	22	0	0	0	48	0
29	45	38	0	0	0	17	0	34	49	0	0	0	17	0
30	42	35	0	0	0	23	0	40	33	0	0	0	27	0
31	36	27	0	0	0	37	0	31	25	0	0	1	43	0
32	34	38	0	0	0	28	0	31	42	0	0	0	27	0
33	28	40	0	0	0	31	1	31	36	0	0	0	33	0
34	40	46	0	0	0	14	0	29	40	0	0	0	31	0
35	38	34	0	0	0	28	0	35	29	0	0	0	36	0
36	60	22	0	0	0	18	0	54	32	0	0	0	14	0
37	40	46	0	0	0	14	0	40	55	0	0	0	5	0
38	61	20	0	0	0	19	0	64	18	0	2	2	14	0
39	24	22	1	3	5	45	0	28	25	5	5	6	31	0
40	40	40	0	0	0	20	0	45	31	0	0	1	23	0
41	37	39	0	0	0	24	0	38	38	0	0	0	24	0
42	33	24	0	0	0	43	0	28	25	0	0	3	44	0
43	23	26	2	3	1	41	4	26	29	2	2	4	36	1
44	43	35	0	0	0	22	0	38	37	0	0	1	24	0
45	51	32	0	0	4	13	0	51	37	0	0	2	9	1
46	36	50	0	0	0	14	0	42	54	0	0	0	4	0
47	34	41	0	0	0	25	0	26	36	0	0	1	37	0
48	38	37	0	0	0	25	0	40	42	0	0	0	18	0
49	21	37	0	0	8	34	0	21	37	1	0	9	32	0
50	31	55	0	0	0	14	0	29	52	0	0	0	19	0

Table B.9: Species 5; ((((A:1,B:1.0):1.0,C:2):1.0,D:3):1.0,E:4.0);

Number	Posterior Probability of Topology Trees							Posterior Probability of Split Trees						
	1	2	3	4	5	6	7	1	2	3	4	5	6	7
1	19	47	0	0	10	19	5	20	42	0	0	13	17	8
2	13	14	8	3	7	55	0	18	15	0	7	8	51	1
3	17	39	2	0	16	26	0	16	35	2	2	19	22	4
4	20	28	0	0	10	24	18	20	39	0	0	5	24	12
5	22	41	1	1	3	29	3	29	29	1	0	7	28	6
6	25	17	1	9	8	21	19	35	4	2	3	8	21	27
7	3	2	19	18	11	19	28	5	7	11	14	5	19	39
8	48	12	3	7	20	10	0	47	14	3	4	16	12	4
9	17	24	5	4	15	32	3	23	21	9	11	14	19	3
10	21	24	0	0	15	40	0	27	26	0	0	12	28	7
11	23	15	3	4	6	48	1	26	16	0	0	13	44	1
12	34	31	0	0	1	31	3	23	29	0	0	24	23	1
13	36	32	0	2	6	17	7	31	27	4	1	5	20	12
14	41	27	0	0	3	29	0	38	30	0	0	8	19	5
15	19	7	1	0	21	36	16	14	5	4	1	18	28	30
16	35	9	0	0	11	32	13	31	13	0	0	10	29	17
17	24	38	0	0	3	33	2	33	25	3	3	5	27	4
18	27	13	0	0	23	32	5	22	10	0	0	24	41	3
19	10	16	3	3	10	40	18	12	16	2	3	10	41	16
20	11	22	3	5	19	24	16	12	17	2	1	18	29	21
21	30	50	0	0	6	14	0	35	47	0	0	4	14	0
22	31	35	0	0	7	24	3	17	37	1	2	3	29	11
23	35	13	0	9	5	22	16	45	29	0	2	0	16	8
24	18	39	0	0	4	39	0	14	36	1	1	4	40	4
25	9	44	0	0	9	38	0	11	44	0	0	8	37	0
26	29	18	1	3	11	26	12	24	13	0	1	28	20	14
27	12	20	9	8	30	17	4	13	20	9	8	25	18	7
28	37	27	3	3	4	23	3	23	33	5	0	17	19	3
29	25	13	7	8	10	22	15	24	11	9	11	19	16	10
30	21	40	0	0	6	31	2	27	33	0	0	3	37	0
31	45	29	0	0	14	12	0	46	30	0	1	13	9	1
32	34	33	0	0	10	17	6	26	32	0	1	7	31	3
33	38	28	4	10	8	8	4	29	32	4	10	8	11	6
34	43	13	4	0	9	15	16	41	16	2	9	6	14	12
35	32	37	0	0	4	27	0	33	41	0	0	5	21	0
36	27	24	6	6	12	15	10	30	23	3	4	14	15	11
37	34	30	0	1	9	25	1	38	30	0	0	8	23	1
38	20	29	0	0	7	44	0	24	20	0	1	10	41	4
39	18	27	1	15	6	17	16	16	20	1	7	9	23	24
40	16	10	0	1	14	44	15	22	17	0	0	8	47	6
41	12	28	1	4	10	44	1	15	19	1	6	16	39	4
42	41	26	0	0	0	32	1	30	31	0	0	6	32	1
43	35	38	0	0	7	18	2	20	38	0	0	9	33	0
44	12	19	2	6	5	42	14	10	20	4	6	9	41	10
45	33	23	0	0	3	39	2	31	18	0	0	8	38	5
46	14	16	7	4	17	33	9	11	28	7	4	22	22	6
47	19	43	0	0	13	17	8	9	48	0	0	18	16	9
48	17	15	0	0	2	62	4	13	13	0	0	16	52	6
49	20	13	11	7	8	20	21	26	14	12	3	7	19	19
50	45	28	0	0	5	20	2	38	18	0	0	15	29	0

Table B.11: Species 6; (((A:1,B:1.0):0.1,C:1.1):0.1,(D:1.1,E:1.1):0.1);

Number	Posterior Probability of Topology Trees							Posterior Probability of Split Trees						
	1	2	3	4	5	6	7	1	2	3	4	5	6	7
1	15	37	7	3	20	13	5	19	20	17	8	10	19	7
2	44	49	0	0	3	4	0	26	46	0	0	28	0	0
3	30	33	4	1	12	16	4	21	43	7	4	18	4	3
4	32	13	0	2	45	8	0	43	32	3	6	14	2	0
5	29	23	2	6	13	13	14	27	21	12	9	11	7	13
6	7	21	5	20	15	17	15	4	11	0	49	13	13	10
7	49	26	0	0	18	7	0	48	20	1	0	23	5	3
8	4	12	20	3	37	21	3	10	11	29	5	31	11	3
9	71	8	4	2	9	4	2	49	12	4	7	22	3	3
10	51	41	0	2	2	4	0	45	35	0	1	5	11	3
11	85	14	0	0	0	1	0	71	23	0	0	1	5	0
12	59	29	0	7	0	5	0	34	39	5	18	2	2	0
13	28	45	2	1	15	2	7	30	34	5	13	15	1	2
14	19	12	8	2	19	33	7	16	24	10	2	20	22	6
15	49	31	0	0	14	6	0	39	36	0	0	23	2	0
16	54	30	3	0	8	3	2	51	30	2	1	10	4	2
17	25	58	0	1	4	12	0	14	55	4	13	8	5	1
18	28	48	0	1	21	2	0	28	24	2	9	32	5	0
19	29	24	1	1	11	12	22	22	26	0	2	22	13	15
20	6	2	44	15	26	3	4	11	7	32	25	23	1	1
21	37	26	2	2	26	6	1	28	41	4	3	12	8	4
22	34	44	0	0	3	19	0	22	35	2	0	16	25	0
23	18	31	12	6	14	13	6	11	35	6	7	30	10	1
24	26	24	3	1	20	23	3	25	38	8	3	19	5	2
25	28	32	0	0	32	7	1	32	30	1	0	30	6	1
26	7	7	2	1	59	16	8	10	6	17	3	46	14	4
27	38	46	0	0	0	14	2	32	41	0	0	3	16	8
28	4	92	0	1	0	3	0	28	59	0	4	1	8	0
29	69	18	0	0	1	12	0	58	16	0	0	10	16	0
30	64	14	0	0	14	5	3	53	8	0	0	32	5	2
31	19	38	1	2	3	32	5	23	35	1	9	3	21	8
32	82	5	0	1	5	7	0	58	15	5	8	8	6	0
33	33	40	0	0	8	19	0	48	44	0	0	5	3	0
34	8	48	12	0	12	12	8	7	29	12	1	35	13	3
35	8	27	2	12	39	3	9	4	18	5	3	66	1	3
36	6	3	13	39	32	3	4	6	9	13	31	31	8	2
37	25	23	0	9	10	18	15	24	19	3	15	25	9	5
38	0	0	63	29	8	0	0	0	0	56	29	14	1	0
39	12	11	12	0	60	4	1	3	20	12	1	57	4	3
40	28	15	10	16	19	11	1	29	12	11	12	33	3	0
41	45	39	0	1	1	12	2	50	30	1	0	3	15	1
42	24	29	7	1	38	0	1	19	25	7	9	27	9	4
43	30	15	1	2	22	21	9	33	14	5	4	19	18	7
44	38	53	0	0	3	6	0	41	45	0	0	6	8	0
45	29	32	1	1	17	19	1	40	27	8	0	18	7	0
46	26	27	9	8	5	10	15	24	35	14	2	12	8	5
47	25	12	1	1	30	26	5	10	14	10	9	24	22	11
48	10	30	1	6	35	13	5	8	37	4	7	24	18	2
49	12	3	23	32	15	13	2	17	4	15	46	10	2	6
50	19	69	2	2	0	8	0	38	40	0	2	0	19	1

Table B.13: Species 7; (((A:1.0,B:1.0):1.0,C:2.0):1.0,(D:2.0,E:2.0):1.0);

Number	Posterior Probability of Topology Trees							Posterior Probability of Split Trees						
	1	2	3	4	5	6	7	1	2	3	4	5	6	7
1	20	15	9	11	7	29	9	11	8	4	16	12	41	8
2	6	2	14	0	25	10	43	3	0	13	3	27	6	48
3	15	14	2	11	20	16	22	12	8	12	10	14	18	26
4	29	28	12	3	12	7	9	36	31	12	0	4	9	8
5	15	15	3	2	15	43	7	19	22	5	5	12	29	8
6	25	13	2	6	14	23	17	20	12	3	2	19	30	14
7	25	17	7	12	14	10	15	24	8	14	18	4	14	18
8	21	30	13	5	0	26	5	21	29	12	4	4	26	4
9	14	28	4	5	2	29	18	17	32	1	1	5	32	12
10	12	13	7	17	4	40	7	12	24	6	11	9	33	5
11	19	14	14	9	7	18	19	21	14	7	7	11	27	13
12	10	8	12	7	15	38	10	9	9	14	7	16	33	12
13	19	11	5	15	14	24	12	11	13	6	17	17	25	11
14	14	22	8	10	12	25	9	9	15	14	5	13	27	17
15	16	8	20	9	14	22	11	12	7	23	13	17	17	11
16	26	17	4	6	6	16	25	17	14	4	5	11	19	30
17	22	10	14	13	11	25	5	19	13	12	22	9	19	6
18	25	24	2	1	5	39	4	24	17	5	3	9	40	2
19	8	17	17	13	23	14	8	16	23	11	11	24	5	10
20	10	16	18	6	18	18	14	17	16	17	4	12	24	10
21	13	13	6	3	20	38	7	29	23	3	2	11	29	3
22	4	2	6	16	34	28	10	6	5	8	12	25	40	4
23	14	10	6	11	22	25	12	11	14	8	10	21	27	9
24	24	9	0	2	10	48	7	23	13	2	2	11	41	8
25	8	11	10	11	11	36	13	7	12	7	6	16	45	7
26	9	11	5	9	18	43	5	11	11	7	11	15	39	6
27	23	20	6	17	3	22	9	24	18	2	22	3	25	6
28	10	10	10	8	12	40	10	13	11	7	6	6	44	13
29	12	20	8	10	19	26	5	8	19	10	8	19	31	5
30	10	26	14	5	5	30	10	10	27	16	4	3	34	6
31	16	16	2	5	4	30	27	16	12	2	9	3	28	30
32	7	20	13	23	13	21	3	12	25	12	19	10	18	4
33	3	3	9	16	39	8	22	2	2	15	7	34	9	31
34	31	9	3	2	9	36	10	23	12	8	5	10	31	11
35	14	29	2	13	5	24	13	14	21	7	7	10	24	17
36	27	17	7	3	1	41	4	31	20	1	5	5	35	3
37	18	23	2	6	16	28	7	17	20	4	7	17	25	10
38	20	14	13	8	8	31	6	13	7	23	7	15	34	1
39	15	8	9	11	10	38	9	16	6	12	16	9	30	11
40	24	21	5	3	4	35	8	26	27	3	5	4	28	7
41	20	18	17	10	11	12	12	14	21	19	7	8	17	14
42	15	10	4	9	35	23	4	14	8	7	11	15	31	14
43	27	14	2	6	12	17	22	21	12	6	6	16	23	16
44	10	13	18	5	8	36	10	11	11	16	9	9	32	12
45	7	7	21	20	8	28	9	11	12	15	16	10	29	7
46	9	8	0	0	24	50	9	17	15	2	5	6	48	7
47	15	9	7	6	25	32	6	15	14	6	6	17	36	6
48	3	2	30	3	7	51	4	4	3	26	3	8	53	3
49	16	14	1	0	9	59	1	19	20	3	5	7	44	2
50	22	32	3	2	11	29	1	20	27	2	2	11	38	0

Table B.15: Species 8; (((A:1,B:1.0):0.1,C:1.1):0.1,(D:0.2,E:0.2):1.0);

Number	Posterior Probability of Topology Trees							Posterior Probability of Split Trees						
	1	2	3	4	5	6	7	1	2	3	4	5	6	7
1	0	0	7	9	73	8	3	3	0	16	9	50	21	1
2	1	1	31	11	12	12	32	1	1	42	17	16	3	20
3	4	11	17	14	18	33	3	8	5	14	13	35	20	5
4	4	11	17	21	17	19	11	4	7	26	21	17	16	9
5	0	0	35	47	3	13	2	2	0	32	39	10	15	2
6	18	12	13	12	30	10	5	7	7	23	11	32	16	4
7	0	0	9	4	75	11	1	1	0	21	7	53	17	1
8	0	0	14	63	12	11	0	0	0	21	45	20	13	1
9	4	5	19	11	19	42	0	11	3	13	17	10	46	0
10	29	24	0	0	22	21	4	33	21	0	2	23	19	2
11	31	12	11	3	16	27	0	17	15	14	4	21	29	0
12	1	1	36	30	17	13	2	0	0	24	41	16	13	6
13	22	18	14	4	5	37	0	13	8	14	15	5	45	0
14	23	8	20	8	3	33	5	12	13	37	6	6	24	2
15	16	14	13	21	12	22	2	16	8	13	29	15	16	3
16	0	0	0	8	58	28	6	3	0	0	17	37	38	5
17	6	8	26	32	11	17	0	8	10	28	22	13	19	0
18	1	0	30	22	20	23	4	0	1	28	28	23	16	4
19	36	15	0	2	14	33	0	26	10	4	5	25	29	1
20	3	34	8	4	20	20	11	7	30	4	6	16	28	9
21	3	12	15	9	15	46	0	4	20	5	6	18	47	0
22	12	19	21	3	32	13	0	17	30	15	2	23	13	0
23	40	24	3	0	5	20	8	24	23	0	2	13	30	8
24	9	2	12	24	14	37	2	10	2	17	25	17	27	2
25	5	14	8	19	12	37	5	3	22	11	10	6	39	9
26	23	4	7	16	17	30	3	20	10	8	10	19	28	5
27	0	0	21	16	53	6	4	0	0	23	26	42	8	1
28	8	2	3	20	14	47	6	8	1	6	17	25	39	4
29	8	21	2	0	16	44	9	19	28	2	1	9	37	4
30	13	15	1	1	8	61	1	4	23	7	3	19	43	1
31	5	2	12	28	23	27	3	1	1	16	13	27	33	9
32	8	4	6	41	35	5	1	8	1	8	36	44	2	1
33	15	18	10	5	26	23	3	14	9	14	9	22	28	4
34	15	22	10	27	5	14	7	4	9	12	35	8	22	10
35	22	15	2	1	6	54	0	12	19	8	4	15	42	0
36	13	18	10	2	20	31	6	3	7	13	14	18	40	5
37	5	8	3	29	18	17	20	17	6	6	34	17	7	13
38	1	1	12	21	53	7	5	0	2	15	15	64	2	2
39	10	10	13	13	41	5	8	9	11	19	11	30	7	13
40	19	21	4	14	15	22	5	11	23	1	23	11	28	3
41	1	1	44	6	29	16	3	1	0	40	0	30	28	1
42	0	3	31	14	38	9	5	4	15	44	7	18	10	2
43	11	5	7	35	22	19	1	12	4	14	20	33	15	2
44	36	20	8	7	6	22	1	30	12	10	12	17	17	2
45	18	15	13	40	7	6	1	8	12	11	52	5	10	2
46	3	0	1	6	61	23	6	6	0	1	3	67	18	5
47	6	1	6	1	74	12	0	8	5	19	4	49	15	0
48	25	21	2	0	24	27	1	20	10	2	3	25	38	2
49	21	4	2	9	10	47	7	17	14	4	27	16	17	5
50	4	9	6	27	5	48	1	11	10	5	17	8	48	1

Table B.17: Species 9; (((A:1,B:1.0):0.1,C:1.1):1.0,(D:2.0,E:2.0):0.1);

Number	Posterior Probability of Topology Trees							Posterior Probability of Split Trees						
	1	2	3	4	5	6	7	1	2	3	4	5	6	7
1	33	34	1	4	5	22	1	32	37	0	1	4	26	0
2	28	48	0	0	1	22	1	26	49	0	2	0	21	2
3	33	46	0	1	0	20	0	48	29	0	1	0	22	0
4	36	33	0	0	0	31	0	42	24	0	0	0	34	0
5	37	56	0	0	0	7	0	29	55	0	0	0	16	0
6	38	60	0	0	0	2	0	37	60	0	0	0	3	0
7	42	27	0	0	1	30	0	33	35	0	0	2	30	0
8	40	38	1	1	0	20	0	38	36	2	3	2	19	0
9	31	34	0	0	0	35	0	31	32	0	0	0	37	0
10	30	16	1	4	5	44	0	23	17	4	10	11	35	0
11	29	22	6	1	9	33	0	17	21	8	5	9	40	0
12	41	48	0	0	0	9	2	41	46	1	0	0	11	1
13	49	42	0	0	0	9	0	47	39	0	1	0	13	0
14	28	34	0	0	1	37	0	36	32	0	0	1	31	0
15	36	30	0	0	0	34	0	33	28	0	0	0	39	0
16	19	46	1	0	2	31	1	13	30	3	3	1	49	1
17	31	54	0	0	1	14	0	24	56	1	0	1	18	0
18	29	23	0	0	1	47	0	30	21	1	0	1	47	0
19	49	20	0	1	0	30	0	48	21	1	0	0	29	1
20	42	23	0	0	0	35	0	43	32	0	0	0	25	0
21	20	5	14	3	15	43	0	15	4	21	11	17	32	0
22	31	50	0	0	1	18	0	29	39	0	2	4	26	0
23	33	27	0	0	1	39	0	31	29	0	0	3	37	0
24	33	38	0	0	0	29	0	33	30	0	0	0	37	0
25	35	56	0	0	2	7	0	35	46	0	0	5	13	1
26	37	27	1	0	1	34	0	38	22	2	1	8	29	0
27	33	46	0	0	0	21	0	39	41	0	0	0	20	0
28	49	31	0	0	0	20	0	46	27	0	1	1	25	0
29	26	50	0	0	0	24	0	22	45	0	0	0	33	0
30	21	35	0	0	0	44	0	28	32	0	0	0	40	0
31	26	56	0	0	1	17	0	30	57	0	0	0	12	1
32	43	36	0	0	0	21	0	48	42	0	0	0	10	0
33	19	45	0	0	0	36	0	29	29	0	0	0	42	0
34	44	44	0	0	0	12	0	35	51	0	0	1	13	0
35	22	27	0	1	2	48	0	19	17	6	3	10	45	0
36	29	29	0	0	0	42	0	32	29	0	0	2	37	0
37	35	33	0	0	1	31	0	22	28	1	0	2	47	0
38	34	18	1	0	2	45	0	37	21	3	3	1	35	0
39	27	46	0	0	0	27	0	32	27	0	0	0	41	0
40	24	35	0	0	1	40	0	23	35	1	1	3	37	0
41	52	22	0	2	1	23	0	51	21	0	2	2	24	0
42	37	29	0	0	0	34	0	33	29	0	0	0	38	0
43	47	40	0	0	0	13	0	50	32	0	0	0	18	0
44	29	43	0	0	3	25	0	21	43	0	0	1	35	0
45	49	40	0	0	0	11	0	49	35	0	0	0	16	0
46	32	33	1	1	2	31	0	35	26	0	3	7	28	1
47	60	27	0	0	3	9	1	66	23	0	0	2	9	0
48	51	37	0	0	0	12	0	44	38	0	0	0	18	0
49	29	40	0	0	0	31	0	27	38	0	1	0	34	0
50	22	24	0	0	2	52	0	14	25	0	3	5	53	0

Table B.19: Species 10; (((A:1,B:1.0):1.0,C:2.0):0.1,(D:2.0,E:2.0):0.1);

Number	Posterior Probability of Topology Trees							Posterior Probability of Split Trees						
	1	2	3	4	5	6	7	1	2	3	4	5	6	7
1	24	16	0	0	38	6	16	20	22	0	0	46	1	11
2	31	34	0	0	4	11	20	46	17	0	0	6	3	28
3	37	57	0	0	2	2	2	33	44	0	0	12	4	7
4	24	42	5	10	5	9	5	35	32	0	7	18	4	4
5	21	45	0	0	11	20	3	28	34	0	0	20	16	2
6	10	16	0	0	42	5	27	11	14	0	1	60	1	13
7	35	11	0	0	29	16	9	22	10	0	0	56	4	8
8	37	54	0	0	3	6	0	33	58	0	0	3	3	3
9	50	17	0	0	13	13	7	45	24	0	0	15	9	7
10	6	49	0	0	24	19	2	6	50	0	0	24	16	4
11	28	22	0	0	22	1	27	18	36	1	0	24	2	19
12	38	46	0	0	11	0	5	24	39	0	0	27	1	9
13	57	23	0	0	19	0	1	37	13	3	2	44	1	0
14	17	12	0	0	57	2	12	13	11	0	0	62	2	12
15	12	48	0	7	11	12	10	13	41	0	3	11	14	18
16	19	17	0	0	33	5	26	26	20	0	0	24	12	18
17	27	26	0	0	34	9	4	31	21	1	0	33	11	3
18	23	32	0	0	38	7	0	26	24	0	0	42	7	1
19	31	25	0	0	21	11	12	26	33	0	0	15	12	14
20	20	29	0	6	11	2	32	21	32	2	8	7	1	29
21	43	23	0	0	27	0	7	20	37	1	2	32	0	8
22	5	8	0	0	60	8	19	7	10	0	1	62	2	18
23	16	19	7	13	16	2	27	38	19	4	4	16	1	18
24	38	35	0	0	22	4	1	50	21	0	0	18	7	4
25	30	20	3	3	12	8	24	32	16	4	5	9	12	22
26	1	12	11	29	19	2	26	9	36	8	19	8	1	19
27	13	61	0	0	15	7	4	7	60	0	0	13	9	11
28	25	51	0	1	5	15	3	28	49	1	1	6	11	4
29	37	25	0	0	33	3	2	43	33	0	0	17	3	4
30	40	9	0	0	36	14	1	25	2	1	0	44	18	10
31	49	27	0	0	9	8	7	36	17	0	0	22	3	22
32	53	26	2	6	9	2	2	70	9	2	5	5	2	7
33	18	15	5	6	43	5	8	12	11	1	3	42	15	16
34	31	21	0	0	33	6	9	21	21	0	1	38	3	16
35	43	20	0	0	31	5	1	62	7	0	0	24	7	0
36	38	15	0	0	26	8	13	39	8	0	1	29	4	19
37	19	35	1	4	32	0	9	18	42	1	3	27	4	5
38	31	19	0	0	21	16	13	31	14	0	0	19	19	17
39	10	66	0	0	14	6	4	10	62	0	0	22	1	5
40	14	68	1	0	8	9	0	29	52	0	0	12	6	1
41	40	25	0	4	18	5	8	40	29	0	0	17	8	6
42	36	28	0	0	21	15	0	23	24	0	0	43	9	1
43	35	17	0	0	47	1	0	30	33	0	0	29	8	0
44	20	52	0	0	12	7	9	18	62	0	0	10	7	3
45	48	34	0	0	8	10	0	46	36	0	0	8	10	0
46	40	48	0	0	8	1	3	39	48	0	0	13	0	0
47	38	37	0	0	16	8	1	40	33	0	0	20	7	0
48	32	18	0	0	45	2	3	29	29	0	1	29	5	7
49	24	43	0	0	9	19	5	22	42	0	0	17	9	10
50	36	36	5	2	9	5	7	26	20	10	12	16	2	14

Table B.21: Species 11; (((A:1,B:1.0):0.1,(D:1.0,E:1.0):0.1):0.1,C:1.2);

Number	Posterior Probability of Topology Trees							Posterior Probability of Split Trees						
	1	2	3	4	5	6	7	1	2	3	4	5	6	7
1	0	0	23	45	23	3	6	2	0	19	41	28	4	6
2	0	0	0	0	99	1	0	2	1	0	0	94	2	1
3	2	1	0	2	86	5	4	10	3	0	6	75	4	2
4	3	1	34	6	44	8	4	3	5	35	15	33	5	4
5	3	2	0	2	85	5	3	9	7	3	3	73	4	1
6	1	3	6	4	65	6	15	1	9	12	5	58	4	11
7	22	6	6	5	48	3	10	21	11	9	6	43	4	6
8	1	0	17	17	57	3	5	1	0	20	21	49	4	4
9	0	0	1	1	93	2	3	0	1	3	5	83	4	4
10	0	0	0	0	92	3	5	0	0	1	2	90	3	4
11	0	0	26	7	61	2	4	0	0	28	7	59	2	4
12	14	4	8	0	53	16	5	15	7	10	2	47	13	6
13	11	62	4	3	9	10	1	16	44	8	8	12	12	0
14	0	3	0	1	94	0	2	4	2	6	4	80	1	3
15	2	1	5	4	79	5	4	5	5	17	13	52	4	4
16	2	4	7	4	71	9	3	6	5	14	16	49	5	5
17	5	1	7	9	60	8	10	5	8	7	23	46	6	5
18	2	0	4	2	80	3	9	4	7	9	6	62	2	10
19	3	2	30	15	36	6	8	4	5	35	15	33	3	5
20	3	8	15	25	43	2	4	6	10	20	25	35	2	2
21	4	8	2	2	69	6	9	6	16	8	3	54	4	9
22	3	4	11	27	45	2	8	3	3	14	28	44	2	6
23	13	6	20	8	35	17	1	9	13	20	9	29	16	4
24	1	0	0	3	94	0	2	3	2	2	7	84	0	2
25	8	12	6	5	60	6	3	12	19	8	6	48	3	4
26	9	7	1	1	65	11	6	18	6	2	3	62	5	4
27	12	11	4	1	51	15	6	18	13	8	9	38	9	5
28	8	6	8	52	15	8	3	8	5	13	44	22	6	1
29	16	15	16	14	27	5	7	15	16	18	24	23	2	2
30	10	5	9	14	51	5	6	15	10	9	21	44	1	0
31	17	25	3	3	26	19	7	23	30	1	3	21	13	9
32	0	0	5	10	75	6	4	0	0	7	17	67	3	6
33	22	3	0	1	67	2	5	22	6	1	2	65	2	2
34	2	0	43	26	18	3	8	3	0	33	29	23	4	8
35	22	7	3	6	29	16	17	31	6	7	7	27	9	13
36	14	22	0	2	51	5	6	13	30	1	3	41	5	7
37	36	16	10	3	16	13	6	29	24	14	5	11	12	5
38	0	0	1	3	89	4	3	2	1	5	6	77	5	4
39	22	12	5	4	46	9	2	21	16	7	8	43	4	1
40	1	0	10	22	55	4	8	2	1	14	22	46	4	11
41	11	7	0	0	74	4	4	11	9	1	4	67	5	3
42	2	2	1	0	94	0	1	1	2	0	0	90	3	4
43	11	19	5	11	49	2	3	16	23	7	10	40	1	3
44	4	28	7	3	38	11	9	14	33	9	7	19	9	9
45	22	25	3	3	33	9	5	24	20	5	5	36	6	4
46	18	18	0	0	57	5	2	18	24	0	0	49	5	4
47	5	74	5	3	10	2	1	9	57	8	10	11	1	4
48	19	13	9	4	37	7	11	24	12	10	7	38	6	3
49	1	0	6	1	89	2	1	3	3	12	8	70	2	2
50	0	0	24	5	60	1	10	0	0	26	14	50	1	9

Table B.23: Species 12; (((A:1,B:1.0):1.0,(D:1.0,E:1.0):1.0):1.0,C:3.0);

Number	Posterior Probability of Topology Trees							Posterior Probability of Split Trees						
	1	2	3	4	5	6	7	1	2	3	4	5	6	7
1	22	31	1	2	13	19	12	26	27	1	1	17	22	6
2	37	38	0	0	1	24	0	31	46	0	0	1	22	0
3	0	0	47	26	6	1	20	0	0	36	23	15	9	17
4	35	37	2	0	12	9	5	38	37	0	0	15	9	1
5	40	35	0	0	3	21	1	47	30	0	0	4	18	1
6	49	37	0	0	0	14	0	46	30	1	0	0	23	0
7	0	0	33	33	0	1	33	0	0	29	37	3	3	28
8	6	16	13	14	12	29	10	4	14	12	12	13	33	12
9	8	12	13	13	16	29	9	5	6	13	10	16	34	16
10	0	0	4	95	0	0	1	3	1	18	69	2	1	6
11	48	49	0	0	0	3	0	47	51	0	0	0	2	0
12	15	11	15	13	5	26	15	10	8	20	13	8	28	13
13	37	24	2	1	17	12	7	37	19	0	0	22	11	11
14	9	10	13	10	18	16	24	8	13	13	10	12	18	26
15	18	17	13	9	11	19	13	21	9	10	13	9	25	13
16	12	18	17	15	11	26	1	13	15	17	17	8	30	0
17	30	30	0	0	8	29	3	24	33	0	0	18	23	2
18	9	13	11	8	11	26	22	11	16	12	5	9	32	15
19	41	37	0	0	2	20	0	44	25	0	0	10	19	2
20	30	31	0	0	1	38	0	32	29	0	0	4	33	2
21	0	0	36	51	0	0	13	0	0	38	47	2	0	13
22	6	12	10	9	11	33	19	7	8	14	9	13	28	21
23	1	0	2	1	67	10	19	5	4	4	2	50	19	16
24	8	13	12	11	15	16	25	9	12	14	16	13	11	25
25	0	0	45	48	0	0	7	0	0	38	54	2	0	6
26	2	2	10	10	45	21	10	4	1	15	11	29	16	24
27	17	14	6	2	22	17	22	7	17	4	4	38	12	18
28	0	1	33	18	5	2	41	0	1	26	24	6	8	35
29	1	2	34	30	13	9	11	0	1	31	37	12	10	9
30	34	30	1	0	29	2	4	33	32	0	0	25	4	6
31	46	46	0	0	0	8	0	58	35	0	0	0	7	0
32	46	47	0	0	0	7	0	40	50	0	0	0	10	0
33	4	10	23	14	16	5	28	13	15	13	14	16	5	24
34	0	0	22	48	1	0	29	2	3	19	35	10	2	29
35	9	12	18	27	7	10	17	8	9	13	21	18	13	18
36	14	18	7	10	13	23	15	13	19	9	12	12	18	17
37	10	10	14	13	17	29	7	9	7	18	11	23	26	6
38	13	10	16	22	7	27	5	6	12	16	16	11	33	6
39	7	13	12	12	14	2	40	6	14	15	18	13	1	33
40	38	30	1	0	9	15	7	32	33	0	2	9	19	5
41	0	0	41	36	5	0	18	1	0	35	36	10	1	17
42	8	4	19	26	5	0	38	4	2	23	28	3	2	37
43	14	14	9	6	26	13	18	11	9	15	6	23	14	22
44	5	2	0	6	59	13	15	7	10	1	2	45	17	18
45	9	6	10	16	9	35	15	9	9	12	12	12	31	15
46	11	9	8	6	11	34	21	9	8	7	7	12	38	19
47	32	27	2	2	17	6	14	40	29	0	1	9	8	13
48	9	13	9	2	14	9	44	11	13	10	3	13	8	42
49	10	9	12	16	19	14	20	11	9	17	18	17	12	16
50	16	9	10	8	20	26	11	15	9	8	13	20	24	11

Table B.25: Species 13; (((A:1,B:1.0):0.1,(D:1.0,E:1.0):0.1):1.0,C:2.1);

Number	Posterior Probability of Topology Trees							Posterior Probability of Split Trees						
	1	2	3	4	5	6	7	1	2	3	4	5	6	7
1	0	0	0	0	97	1	2	0	1	2	1	87	6	3
2	1	0	0	0	98	1	0	1	2	2	1	89	1	4
3	12	47	0	0	34	4	3	14	36	2	2	37	6	3
4	0	0	0	0	99	1	0	2	0	2	2	89	1	4
5	9	31	0	0	52	6	2	12	26	1	0	50	5	6
6	1	8	8	3	66	9	5	5	10	11	13	53	4	4
7	3	2	1	1	89	4	0	7	9	3	2	73	2	4
8	1	0	1	0	94	2	2	2	7	2	1	77	6	5
9	1	1	0	1	96	0	1	1	2	8	4	80	3	2
10	0	0	2	2	96	0	0	0	1	6	7	83	1	2
11	0	1	2	1	92	4	0	4	2	6	6	73	8	0
12	0	2	0	0	97	0	1	2	8	0	1	83	1	5
13	1	0	0	0	98	1	0	4	3	2	2	85	2	2
14	0	1	7	3	82	2	5	3	3	17	6	65	2	3
15	3	1	31	19	33	1	12	2	3	35	26	28	0	6
16	0	1	16	15	59	3	6	0	2	19	19	48	4	8
17	2	0	1	0	91	2	4	4	4	5	6	67	7	7
18	0	0	1	3	90	4	2	4	1	3	6	80	3	3
19	17	7	5	4	50	13	4	28	12	4	10	36	7	3
20	0	2	0	0	96	2	0	3	5	0	1	87	1	2
21	37	11	5	3	36	6	2	33	15	5	2	39	3	2
22	37	28	0	0	24	10	1	31	33	3	0	25	6	2
23	0	0	0	0	99	1	0	2	0	1	1	93	1	1
24	0	12	4	2	76	4	2	6	15	10	6	58	3	2
25	1	3	2	0	81	4	9	6	2	5	6	71	2	8
26	22	1	13	4	41	11	8	16	5	17	13	32	10	7
27	3	0	0	0	94	2	1	10	4	0	1	78	4	3
28	3	0	3	3	86	2	3	6	4	8	12	64	2	4
29	0	0	0	0	100	0	0	4	1	2	2	89	2	0
30	6	9	1	0	73	10	1	13	17	1	3	58	5	3
31	0	6	0	2	87	5	0	11	16	3	2	62	3	3
32	1	1	3	16	74	3	2	4	2	9	25	54	2	4
33	0	0	0	0	95	2	3	3	2	4	2	81	4	4
34	1	0	4	4	85	2	4	6	3	7	9	70	3	2
35	1	2	1	0	91	5	0	6	5	3	1	72	11	2
36	12	7	0	0	70	5	6	17	10	2	0	63	4	4
37	0	0	0	0	100	0	0	1	5	1	0	89	2	2
38	1	0	4	1	88	5	1	2	0	6	2	78	8	4
39	3	2	0	0	91	3	1	16	3	3	1	70	4	3
40	5	4	0	0	83	1	7	11	9	6	4	59	0	11
41	1	1	6	25	63	2	2	4	4	6	27	53	2	4
42	5	2	0	4	83	4	2	7	8	2	3	73	5	2
43	0	0	1	0	98	1	0	1	5	5	0	84	4	1
44	1	1	4	0	85	9	0	3	5	9	5	72	6	0
45	1	2	1	1	92	2	1	3	14	6	4	69	2	2
46	2	1	1	5	90	0	1	5	3	3	6	75	4	4
47	0	0	4	3	89	1	3	0	0	7	4	85	1	3
48	3	3	4	3	77	1	9	7	5	8	8	60	4	8
49	0	2	29	2	57	5	5	1	4	30	3	55	3	2
50	10	15	1	3	60	3	8	17	24	4	1	47	2	5

Table B.27: Species 14; (((A:1,B:1.0):0.1,(D:0.1,E:0.1):1.0):0.1,C:1.2);

Number	Posterior Probability of Topology Trees							Posterior Probability of Split Trees						
	1	2	3	4	5	6	7	1	2	3	4	5	6	7
1	0	0	30	41	12	13	4	0	0	22	38	26	11	3
2	0	0	18	43	35	1	3	0	0	19	31	44	2	4
3	0	0	29	28	34	5	3	2	1	24	27	32	10	3
4	0	0	50	25	15	7	3	0	0	42	26	22	8	2
5	1	0	41	21	21	11	5	0	0	31	25	25	15	4
6	1	4	12	27	30	17	9	2	2	15	22	33	18	8
7	0	0	16	8	63	1	12	0	0	16	15	55	3	10
8	4	1	17	16	35	21	6	8	3	15	20	38	14	2
9	4	2	20	16	35	19	4	5	3	19	18	32	19	4
10	0	0	35	23	29	7	6	0	0	23	24	40	8	5
11	0	0	15	15	48	6	16	0	0	23	29	27	9	12
12	1	3	46	20	14	12	4	0	0	45	29	10	10	6
13	0	0	35	21	29	8	7	0	0	29	20	39	8	4
14	0	3	13	15	53	12	4	0	0	20	15	53	9	3
15	2	7	6	11	44	21	9	15	12	3	11	27	26	6
16	2	0	22	11	52	10	3	0	1	23	13	49	12	2
17	3	5	24	18	24	15	11	3	3	22	19	25	21	7
18	6	6	13	11	30	33	1	6	10	17	10	27	26	4
19	2	6	22	35	17	11	7	5	6	27	32	14	12	4
20	12	9	20	15	25	17	2	5	7	27	27	18	10	6
21	8	8	22	32	12	17	1	5	3	21	29	24	14	4
22	8	8	35	15	12	20	2	5	6	31	25	9	22	2
23	2	4	20	24	26	22	2	6	6	17	20	28	21	2
24	10	15	18	13	14	29	1	11	6	24	13	21	25	0
25	0	0	31	36	28	3	2	0	0	26	32	33	5	4
26	0	0	33	41	17	1	8	0	0	38	35	18	3	6
27	2	0	43	22	17	12	4	0	1	44	29	18	5	3
28	3	3	12	10	48	23	1	1	0	15	22	34	23	5
29	7	7	7	6	41	26	6	7	8	11	13	31	24	6
30	9	5	32	25	10	14	5	9	6	28	28	9	16	4
31	0	0	41	30	17	5	7	0	0	36	24	26	9	5
32	0	0	33	18	34	9	6	0	0	32	30	31	7	0
33	10	16	2	7	34	29	2	7	16	13	12	22	24	5
34	0	0	25	52	15	4	4	0	0	30	49	14	4	3
35	9	11	22	14	22	19	3	10	11	22	17	18	17	5
36	0	0	16	30	48	4	2	0	0	17	27	49	4	3
37	0	0	49	15	28	7	1	2	0	37	17	34	9	1
38	0	0	8	16	62	7	7	0	0	16	30	41	6	7
39	0	0	27	27	34	7	5	0	0	30	25	35	6	4
40	8	8	16	14	32	18	4	2	4	18	20	31	20	5
41	3	3	15	2	55	19	3	2	2	21	6	52	15	2
42	0	0	21	14	46	9	10	0	0	26	25	39	3	7
43	3	5	19	28	18	23	4	1	7	17	30	18	24	3
44	0	0	25	15	44	11	5	0	0	29	18	41	3	9
45	0	0	21	25	39	9	6	0	0	21	21	42	11	4
46	3	3	31	17	17	18	11	1	2	25	21	27	16	8
47	0	0	43	25	28	2	2	0	0	38	29	24	7	2
48	0	0	46	33	10	2	9	0	0	38	36	17	2	7
49	0	0	3	8	76	9	4	0	0	12	13	64	9	2
50	0	0	11	11	58	15	5	2	0	20	15	41	14	8

B.2

Table B.29: output of Species 2 With DNA Sequences

Number	Posterior Probability of Topology Trees							Posterior Probability of Split Trees							Note
	1	2	3	4	5	6	7	1	2	3	4	5	6	7	
1	66	18	0	4	7	3	2	45	26	3	7	13	3	2	*
2	67	16	3	1	10	1	2	45	22	9	7	14	0	3	
3	69	11	2	0	5	10	3	51	15	3	7	10	9	5	
4	60	26	1	1	5	4	3	56	20	4	2	11	2	5	
5	63	20	0	2	4	9	2	52	20	5	3	5	12	3	
6	56	22	2	5	9	6	0	48	19	5	7	14	6	1	
7	57	20	3	5	9	3	3	52	21	5	5	13	3	1	
8	46	29	4	0	8	8	5	50	18	5	6	10	8	3	
9	56	29	0	0	8	2	5	46	26	2	3	13	3	6	*
10	61	24	3	1	7	4	0	50	25	8	2	10	5	0	

Table B.31: output of Species 8 With DNA Sequences

Number	Posterior Probability of Topology Trees							Posterior Probability of Split Trees						
	1	2	3	4	5	6	7	1	2	3	4	5	6	7
1	7	9	18	23	17	22	4	7	9	22	17	25	17	3
2	5	4	24	25	20	17	5	6	3	29	16	24	18	4
3	11	6	25	15	17	21	5	9	8	22	15	22	20	4
4	5	6	20	19	22	23	5	7	7	17	21	21	24	3
5	3	6	20	21	23	23	4	4	11	16	28	22	15	4
6	6	6	26	17	17	21	7	8	5	26	26	19	11	5
7	3	9	24	23	16	20	5	5	2	23	16	23	22	9
8	9	9	17	17	21	26	1	9	12	4	19	28	27	1
9	7	8	21	24	24	13	3	9	11	16	24	16	19	5
10	8	7	19	20	19	22	5	5	7	22	14	22	25	5

Table B.33: output of Species 13 With DNA Sequences

Number	Posterior Probability of Topology Trees							Posterior Probability of Split Trees						
	1	2	3	4	5	6	7	1	2	3	4	5	6	7
1	9	11	4	4	61	3	8	7	14	10	9	53	3	4
2	5	2	7	9	62	9	6	13	7	5	8	53	8	6
3	12	8	5	6	61	5	3	13	7	8	11	51	5	5
4	10	4	9	9	63	1	4	10	7	12	14	48	5	4
5	9	4	3	5	68	4	7	13	4	10	8	58	3	4
6	9	13	9	9	44	7	9	14	12	10	9	41	5	9
7	5	3	10	14	59	6	3	4	6	10	9	58	7	6
8	14	8	2	8	58	4	6	13	8	6	12	51	4	6
9	3	4	1	4	76	5	7	9	14	4	6	53	5	9
10	5	6	9	10	61	9	0	14	12	11	10	45	6	2

B.3

Table B.35: Output of Caterpillar Tree for Eight Taxa

	posterior of Topology	RF dist	AB	ABC	ABCD	ABCDE	ABCDEF	ABCDEFG	GH
1	26	2	97	95	92	83	54	26	51
2	26	2	100	100	98	88	63	26	49
3	20	2	97	94	91	81	54	20	52
4	29	2	99	99	96	83	54	29	55
5	16	2	97	95	90	80	51	16	59
6	26	2	99	98	97	89	60	26	50
7	28	2	96	92	90	77	58	28	50
8	20	2	100	97	95	85	55	20	58
9	28	2	97	97	93	79	58	28	49
10	23	2	99	98	95	82	51	23	53
11	28	2	100	99	98	89	65	28	43
12	23	2	97	92	90	79	47	23	58
13	34	2	98	96	95	87	63	34	46
14	27	2	98	97	95	78	55	27	48
15	31	2	97	93	90	81	52	31	51
16	19	2	98	98	96	86	52	19	59
17	24	2	100	98	94	78	50	24	54
18	17	2	95	93	87	77	44	17	64
19	22	2	99	94	90	77	59	22	50
20	15	2	100	98	92	82	45	15	62
21	21	2	95	92	86	78	48	21	58
22	18	2	100	100	96	86	58	18	50
23	25	2	100	98	94	85	54	25	49
24	29	2	96	93	90	80	59	29	46
25	27	2	97	92	85	76	46	27	57
26	19	2	97	95	88	76	43	19	60
27	22	2	99	97	94	86	56	22	52
28	20	2	100	98	91	84	46	20	56
29	30	2	100	99	98	86	58	30	47
30	24	2	100	99	94	83	59	24	45
31	32	2	100	99	96	86	63	32	45
32	24	2	96	95	92	85	55	24	52
33	25	2	99	99	98	87	60	25	49
34	30	2	97	95	93	82	59	30	46
35	19	2	94	93	87	78	48	19	59
36	32	2	98	95	93	82	66	32	39
37	30	2	99	99	98	90	60	31	49
38	20	2	99	98	97	86	58	20	49
39	16	2	99	97	94	83	47	16	60
40	25	2	98	95	92	78	54	25	51
41	25	2	99	99	95	88	61	25	46
42	19	2	96	94	91	87	56	19	53
43	28	2	98	96	93	85	58	28	44
44	24	2	99	99	94	80	55	24	52
45	37	2	99	97	95	87	67	37	38
46	29	2	100	99	95	82	55	29	51
47	16	2	98	95	94	78	44	16	59
48	27	2	97	95	93	77	60	27	44
49	19	2	100	97	94	81	45	19	60
50	29	2	99	97	90	81	49	29	56

Table B.37: Output of Balance Tree for Eight Taxa

	posterior of Topology	RF dist	AB	CD	ABCD	EF	GH	EFGH
1	23	0	79	78	48	88	93	75
2	21	0	82	75	46	94	90	75
3	29	0	86	87	58	87	84	71
4	27	0	79	88	50	90	89	77
5	29	0	83	82	51	95	88	78
6	22	0	85	77	49	83	90	73
7	31	0	83	82	54	92	87	77
8	24	0	88	77	59	86	85	65
9	29	0	86	83	59	87	88	70
10	17	0	78	79	48	89	81	69
11	29	0	87	79	57	89	88	72
12	23	0	88	79	57	81	92	66
13	15	0	80	87	59	80	83	56
14	23	0	85	79	59	88	85	64
15	29	0	83	83	57	86	91	72
16	28	0	82	84	55	87	89	73
17	19	0	84	76	50	84	91	69
18	16	0	77	78	46	86	90	70
19	22	0	82	80	50	93	84	72
20	25	0	86	85	61	84	84	64
21	19	0	84	79	53	83	87	66
22	30	0	92	77	62	83	86	68
23	31	0	82	82	52	92	88	79
24	28	0	93	85	66	81	85	62
25	15	0	80	77	44	83	91	71
26	20	0	81	85	59	84	83	61
27	21	0	80	86	54	80	91	67
28	25	0	80	77	46	87	94	79
29	29	0	78	87	58	90	89	71
30	20	0	87	80	59	85	80	61
31	23	0	86	83	60	85	83	63
32	17	0	86	77	53	86	85	64
33	28	0	79	88	59	86	93	69
34	31	0	82	83	55	92	91	76
35	26	0	87	83	58	88	88	68
36	22	0	76	87	55	90	81	67
37	24	0	87	84	54	89	86	70
38	26	0	80	80	52	88	91	74
39	26	0	84	84	55	90	83	71
40	28	0	78	80	51	94	83	77
41	32	0	86	88	67	89	89	65
42	27	0	89	82	58	80	92	69
43	20	0	82	82	55	89	82	65
44	25	0	84	84	53	85	91	72
45	21	0	84	73	51	85	90	70
46	22	0	86	85	63	80	86	59
47	28	0	80	82	57	88	87	71
48	31	0	84	87	65	86	84	66
49	30	0	85	89	68	83	87	62
50	28	0	89	85	65	89	81	63

Appendix C

Output for MLE and Bootstrapping method

C.1

Table C.1: Output of MLE and Bootstrapping For Species 1

#	X	Y	Z	Likelihood	Tree	Bootstrapping Trees						
						1	2	3	4	5	6	7
1	92	91	79	-263.7466	1	37	2	1	3	0	4	3
2	91	95	92	-265.7397	6	11	2	5	2	2	18	10
3	95	90	91	-265.6416	1	31	19	0	0	0	0	0
4	86	88	1	-256.4748	1	47	3	0	0	0	0	0
5	94	90	1	-260.1342	1	49	0	0	0	0	1	0
6	91	93	85	-264.4548	1	37	6	1	1	0	0	5
7	79	99	75	-259.912	1	40	2	0	0	4	2	2
8	92	92	92	-264.5097	6	12	2	0	0	13	23	0
9	93	78	97	-256.4972	6	17	7	0	0	0	26	0
10	98	86	81	-262.8554	1	36	8	0	0	0	6	0
11	92	88	92	-261.9833	6	27	0	0	0	6	17	0
12	85	89	88	-260.3466	1	32	2	0	0	5	11	0
13	86	90	73	-259.4897	1	41	8	0	0	0	0	1
14	86	91	95	-263.4135	1	26	8	0	0	3	13	0
15	99	90	97	-267.3793	7	1	1	19	8	0	1	20
16	94	89	72	-262.5622	2	7	42	0	0	0	1	0
17	95	90	94	-265.3021	6	3	19	0	2	1	23	2
18	91	89	1	-259.2106	1	45	0	0	0	0	5	0
19	90	94	76	-264.0921	1	43	5	0	0	1	1	0
20	98	94	97	-269.2104	6	14	8	4	2	5	15	2
21	85	92	90	-262.2867	1	30	16	0	1	0	3	0
22	93	86	1	-257.5151	1	47	3	0	0	0	0	0
23	85	92	87	-261.5141	2	5	41	0	0	0	3	1
24	93	89	98	-265.7362	1	20	13	0	1	3	13	0
25	94	91	96	-266.2565	6	8	13	5	0	7	17	0
26	83	95	92	-261.7519	6	5	5	0	0	7	23	10
27	98	88	95	-265.7439	1	27	14	0	1	0	8	0
28	98	86	73	-262.0745	1	42	6	0	0	0	2	0
29	88	93	97	-265.2051	6	13	7	0	0	16	14	0
30	94	95	1	-264.023	1	48	2	0	0	0	0	0
31	94	90	93	-264.3425	6	12	3	0	2	6	27	0
32	93	90	82	-263.0878	2	31	19	0	0	0	0	0
33	97	91	1	-261.2138	1	46	4	0	0	0	0	0
34	87	90	77	-260.6802	2	3	37	0	0	2	8	0
35	86	96	86	-263.7647	1	36	1	0	1	1	6	5
36	88	96	1	-261.2854	1	50	0	0	0	0	0	0
37	100	88	87	-262.4688	6	12	0	16	0	1	21	0
38	96	86	93	-264.3796	1	27	4	0	0	0	19	0
39	97	93	81	-260.9357	7	2	0	4	0	12	0	32
40	96	79	92	-258.9226	1	28	11	0	0	0	11	0
41	93	96	81	-265.7874	1	41	3	0	1	0	0	5
42	96	95	1	-263.4693	1	45	0	0	0	1	3	1
43	95	89	96	-265.9623	1	24	11	0	0	1	13	1
44	92	84	90	-261.3337	1	32	2	0	0	0	16	0
45	94	98	83	-263.8955	7	6	4	0	2	8	11	19
46	81	96	93	-262.1708	1	27	5	0	0	2	8	8
47	86	89	92	-261.4837	1	26	18	0	0	0	6	0
48	91	95	92	-267.1982	1	28	0	6	2	1	3	10
49	93	95	93	-267.5752	1	34	5	0	1	6	4	0
50	82	90	97	-260.8592	2	14	20	0	1	1	13	1

Table C.3: Output of MLE and Bootstrapping For Species 2

#	X	Y	Z	Likelihood	Tree	Bootstrapping Trees						
						1	2	3	4	5	6	7
1	90	90	1	-254.6614	1	49	1	0	0	0	0	0
2	100	90	1	-263.093	1	43	6	0	0	0	1	0
3	86	93	83	-262.6607	1	27	0	0	0	5	13	5
4	78	91	74	-256.0214	1	41	9	0	0	0	0	0
5	81	87	1	-250.3972	1	50	0	0	0	0	0	0
6	94	88	1	-260.2149	1	47	0	0	0	0	3	0
7	87	94	1	-257.7319	1	49	1	0	0	0	0	0
8	98	87	94	-265.2887	1	30	7	0	0	0	13	0
9	93	97	1	-261.919	1	47	3	0	0	0	0	0
10	85	84	1	-252.6745	1	40	9	0	0	0	0	1
11	98	92	1	-260.0364	1	49	0	0	0	0	1	0
12	100	89	64	-263.53	1	41	9	0	0	0	0	0
13	89	86	85	-260.3092	1	41	7	0	0	0	2	0
14	87	89	96	-262.2509	6	19	2	0	0	5	23	1
15	90	80	1	-251.5361	1	44	6	0	0	0	0	0
16	94	87	1	-257.4073	1	48	2	0	0	0	0	0
17	88	89	1	-255.7076	1	47	1	0	0	1	1	0
18	92	87	85	-262.0107	1	39	8	0	0	0	3	0
19	100	85	73	-262.1074	1	41	7	0	0	0	2	0
20	96	88	1	-261.4295	1	46	4	0	0	0	0	0
21	92	84	1	-254.0752	1	42	8	0	0	0	0	0
22	86	87	1	-255.8913	1	49	1	0	0	0	0	0
23	92	84	1	-254.4022	1	37	13	0	0	0	0	0
24	85	94	95	-263.6747	6	16	6	0	0	1	16	11
25	94	92	75	-264.2569	1	43	4	1	0	0	2	0
26	91	93	98	-266.5504	6	10	17	13	0	0	5	5
27	88	93	80	-263.3019	1	39	3	1	0	3	4	0
28	93	94	85	-265.9219	1	28	22	0	0	0	0	0
29	86	82	84	-255.7079	1	32	14	0	0	0	4	0
30	87	87	1	-256.8296	1	46	2	0	0	0	2	0
31	79	96	86	-259.8175	1	34	4	0	0	5	4	3
32	93	91	1	-262.567	1	48	1	0	0	0	1	0
33	83	87	1	-254.4576	1	48	1	0	0	0	1	0
34	96	96	99	-269.7092	6	7	14	2	4	10	8	5
35	87	89	95	-262.2226	2	22	19	0	0	8	1	0
36	87	93	76	-261.8802	1	38	4	0	0	7	1	0
37	91	91	88	-264.1069	1	33	6	0	0	0	8	3
38	93	87	93	-263.1204	1	20	24	0	0	0	6	0
39	85	87	1	-255.4122	1	43	7	0	0	0	0	0
40	87	88	75	-259.4678	1	40	1	2	0	0	5	2
41	94	93	1	-259.0934	1	49	0	0	0	0	1	0
42	78	90	93	-257.706	1	30	5	0	0	1	13	1
43	91	92	68	-262.3838	2	30	20	0	0	0	0	0
44	91	88	90	-262.7418	2	24	22	0	0	0	4	0
45	86	93	97	-264.0182	6	14	10	2	0	4	14	6
46	91	90	90	-264.3417	1	23	1	0	0	3	23	0
47	95	89	94	-265.51	1	27	19	0	0	2	2	0
48	94	96	1	-264.6761	1	48	2	0	0	0	0	0
49	76	90	89	-251.9163	6	19	1	0	0	6	23	1
50	90	87	1	-257.362	1	49	0	0	0	0	1	0

Table C.5: Output of MLE and Bootstrapping For Species 3

#	X	Y	Z	Likelihood	Tree	Bootstrapping Trees						
						1	2	3	4	5	6	7
1	35	100	85	-205.1585	1	31	4	0	0	2	1	12
2	37	93	1	-202.7226	1	42	7	0	0	1	0	0
3	38	93	88	-209.2292	2	4	29	0	0	7	2	8
4	44	97	1	-215.7373	2	0	50	0	0	0	0	0
5	37	88	1	-204.2905	1	27	0	0	0	14	6	3
6	34	75	49	-189.5318	2	9	31	0	0	3	7	0
7	26	88	1	-182.3551	2	26	20	0	0	0	1	3
8	91	99	32	-199.7307	7	0	5	0	0	21	7	17
9	30	92	85	-194.3162	1	26	4	0	0	8	4	8
10	47	83	1	-213.4206	1	45	4	0	0	0	0	1
11	29	92	92	-191.9621	1	24	11	0	0	0	2	13
12	30	87	1	-189.9398	2	3	35	0	0	0	3	9
13	41	88	1	-208.9313	2	16	32	0	0	2	0	0
14	35	96	90	-203.4602	6	15	1	0	0	11	9	14
15	44	87	1	-214.8666	2	16	32	0	0	0	1	1
16	32	75	1	-185.9808	1	31	17	0	0	1	1	0
17	39	86	75	-206.3336	1	35	5	0	0	3	5	2
18	40	82	78	-204.7212	1	33	9	0	0	3	5	0
19	26	90	1	-181.1052	1	46	3	0	0	0	0	1
20	31	92	98	-197.0514	6	7	11	0	0	29	3	0
21	96	94	31	-197.7338	7	0	1	0	0	19	0	30
22	31	92	1	-192.8672	1	37	0	0	0	9	2	2
23	39	86	81	-206.513	1	26	16	0	0	1	6	1
24	34	92	1	-198.0926	1	42	5	0	0	0	0	3
25	46	86	1	-216.091	1	32	18	0	0	0	0	0
26	35	90	1	-199.8477	1	39	7	0	0	0	1	3
27	32	93	93	-199.8722	2	3	18	0	0	2	6	21
28	28	96	84	-192.269	1	25	0	0	0	0	1	24
29	46	93	1	-219.5174	1	32	0	0	0	14	4	0
30	45	96	1	-219.3294	2	26	21	0	0	3	0	0
31	40	90	93	-213.2978	5	16	2	0	0	24	4	4
32	36	95	77	-206.5379	1	24	1	0	0	23	2	0
33	36	86	1	-199.6063	2	10	33	0	0	1	4	2
34	35	91	98	-204.0894	6	17	9	0	0	0	5	19
35	40	85	100	-209.4639	1	20	16	0	0	2	7	5
36	36	94	1	-203.7653	1	39	4	0	0	3	0	4
37	37	94	80	-206.7632	1	21	3	0	0	17	5	4
38	32	86	1	-191.2577	1	47	3	0	0	0	0	0
39	28	99	1	-191.4961	5	18	0	0	0	32	0	0
40	38	100	77	-211.6561	1	34	5	0	0	10	0	1
41	40	87	93	-209.7615	2	9	22	0	0	5	14	0
42	33	84	1	-193.5596	2	24	24	0	0	0	2	0
43	31	97	1	-194.9776	1	45	0	0	0	0	1	4
44	34	94	87	-202.1541	1	25	7	0	0	14	2	2
45	36	91	95	-206.9207	5	0	4	0	0	22	2	22
46	49	96	1	-224.0915	1	44	0	0	0	1	2	3
47	32	90	1	-194.9674	2	23	27	0	0	0	0	0
48	37	86	1	-200.8787	1	46	2	0	0	0	2	0
49	29	91	88	-191.6671	1	13	13	0	0	13	6	5
50	37	86	1	-199.5729	2	0	45	0	0	2	1	2

Table C.7: Output of MLE and Bootstrapping For Species 4

#	X	Y	Z	Likelihood	Tree	Bootstrapping Trees						
						1	2	3	4	5	6	7
1	100	41	73	-214.3131	1	20	6	0	0	0	24	0
2	93	39	85	-208.6711	1	26	15	0	0	0	9	0
3	99	34	1	-200.5509	2	5	36	0	0	0	9	0
4	94	28	1	-188.9095	2	22	28	0	0	0	0	0
5	88	34	1	-197.5486	2	1	44	0	0	0	5	0
6	100	39	85	-202.2107	6	7	20	0	0	0	23	0
7	79	44	1	-206.4824	1	40	7	0	0	0	3	0
8	95	47	77	-207.4085	6	8	4	0	0	0	38	0
9	95	37	1	-204.1888	1	29	21	0	0	0	0	0
10	85	33	1	-193.6929	1	36	14	0	0	0	0	0
11	68	41	1	-195.6203	2	10	28	0	0	0	12	0
12	78	33	1	-189.6613	1	50	0	0	0	0	0	0
13	94	31	1	-195.6773	1	21	14	0	0	0	15	0
14	80	50	78	-205.2178	6	11	6	0	0	0	33	0
15	81	35	1	-197.1403	1	20	15	0	0	0	15	0
16	74	59	65	-201.9125	6	3	5	0	0	0	42	0
17	79	40	1	-201.7456	2	9	37	0	0	0	4	0
18	96	39	88	-203.532	6	14	14	0	0	0	22	0
19	92	39	1	-206.7669	1	38	10	0	0	0	2	0
20	97	45	1	-216.526	2	8	40	0	0	0	2	0
21	86	41	82	-198.777	6	21	6	0	0	0	23	0
22	96	28	1	-189.1027	2	15	35	0	0	0	0	0
23	86	30	1	-188.8676	1	47	3	0	0	0	0	0
24	93	50	65	-200.9271	6	5	1	0	0	0	44	0
25	93	57	59	-204.6523	6	4	0	0	0	0	46	0
26	78	43	1	-205.4896	1	29	14	0	0	0	7	0
27	79	46	93	-208.3953	6	10	15	0	0	0	25	0
28	89	34	1	-197.4654	2	6	44	0	0	0	0	0
29	91	37	1	-204.2679	1	36	1	0	0	0	13	0
30	89	39	93	-203.9742	6	11	13	0	0	0	26	0
31	88	49	69	-201.5585	6	0	19	0	0	0	31	0
32	90	38	1	-205.4976	1	22	10	0	0	0	18	0
33	97	29	1	-192.7393	2	4	27	0	0	0	19	0
34	94	33	1	-197.0325	2	11	39	0	0	0	0	0
35	94	36	1	-202.5662	1	49	1	0	0	0	0	0
36	79	34	1	-193.3268	2	15	35	0	0	0	0	0
37	93	28	1	-188.8556	1	26	24	0	0	0	0	0
38	86	42	82	-200.3203	6	27	1	0	0	0	22	0
39	100	44	90	-213.1033	6	19	9	0	0	0	22	0
40	83	32	1	-193.2311	1	18	13	0	0	0	19	0
41	96	37	52	-206.3022	1	27	5	0	0	0	18	0
42	88	41	1	-208.7513	1	29	8	0	0	0	13	0
43	82	52	62	-196.6181	6	0	8	0	0	0	42	0
44	86	40	1	-205.1862	2	5	42	0	0	0	3	0
45	85	40	92	-202.8474	6	6	20	0	0	0	24	0
46	85	43	1	-210.4822	1	36	5	0	0	0	9	0
47	84	37	1	-200.963	1	26	0	0	0	0	24	0
48	100	44	87	-211.8161	6	7	19	0	0	0	24	0
49	99	46	1	-219.3059	1	25	22	0	0	0	3	0
50	75	53	69	-198.5872	6	0	9	0	0	0	41	0

Table C.9: Output of MLE and Bootstrapping For Species 5

						Bootstrapping Trees						
#	X	Y	Z	Likelihood	Tree	1	2	3	4	5	6	7
1	39	41	1	-161.8227	1	38	12	0	0	0	0	0
2	43	46	58	-149.1948	6	0	19	0	0	0	31	0
3	42	61	57	-162.8521	6	6	5	0	0	4	35	0
4	40	40	85	-163.8133	1	18	12	0	0	0	20	0
5	36	34	1	-148.5189	1	31	19	0	0	0	0	0
6	36	45	86	-157.1396	6	3	26	1	0	0	18	2
7	33	73	53	-155.6577	6	3	10	0	0	0	27	10
8	27	40	1	-143.2721	1	39	11	0	0	0	0	0
9	48	34	1	-165.8834	1	40	10	0	0	0	0	0
10	34	34	1	-146.5321	1	26	24	0	0	0	0	0
11	33	28	1	-135.4962	2	18	32	0	0	0	0	0
12	32	52	65	-144.6683	6	8	6	0	0	0	36	0
13	27	33	1	-132.6781	1	33	17	0	0	0	0	0
14	32	42	1	-154.5658	1	25	14	0	0	0	11	0
15	39	68	46	-153.958	6	3	8	0	0	0	34	5
16	33	29	1	-137.5078	1	47	3	0	0	0	0	0
17	42	34	1	-158.2126	1	37	13	0	0	0	0	0
18	39	37	1	-157.352	1	31	19	0	0	0	0	0
19	27	25	1	-121.4909	2	10	38	0	2	0	0	0
20	26	64	64	-145.8821	6	0	12	0	1	0	35	2
21	40	49	1	-175.1207	1	36	9	0	0	0	5	0
22	44	31	1	-156.1739	1	46	4	0	0	0	0	0
23	41	48	1	-174.8827	2	13	37	0	0	0	0	0
24	37	46	1	-167.4578	2	3	35	0	0	0	12	0
25	42	36	1	-160.9072	2	4	46	0	0	0	0	0
26	39	76	46	-159.6467	6	0	1	0	0	0	41	8
27	30	38	1	-147.1071	1	38	12	0	0	0	0	0
28	42	41	89	-162.2655	6	12	13	0	0	0	25	0
29	29	64	59	-147.1291	6	10	0	0	0	0	38	2
30	37	35	1	-151.8512	1	39	11	0	0	0	0	0
31	33	41	1	-154.7041	1	25	25	0	0	0	0	0
32	34	41	1	-157.2195	1	33	17	0	0	0	0	0
33	31	26	1	-129.9341	2	20	30	0	0	0	0	0
34	31	41	1	-152.8455	1	28	0	0	0	0	22	0
35	39	38	1	-158.5439	1	44	6	0	0	0	0	0
36	30	36	1	-144.5014	1	31	18	0	1	0	0	0
37	40	31	1	-150.8097	1	34	16	0	0	0	0	0
38	36	53	77	-161.5516	6	13	9	0	0	0	28	0
39	37	46	70	-150.124	6	18	12	0	0	0	20	0
40	33	34	1	-145.2755	1	38	12	0	0	0	0	0
41	40	44	1	-169.5901	2	16	24	0	0	0	10	0
42	34	34	1	-144.2348	2	0	32	0	0	0	18	0
43	38	42	1	-163.0057	1	40	10	0	0	0	0	0
44	31	53	53	-134.5864	6	16	1	1	0	0	32	0
45	52	35	95	-171.7842	1	20	14	0	0	0	16	0
46	36	69	48	-152.1895	6	0	11	0	0	8	29	2
47	39	43	1	-166.5892	1	24	11	0	0	0	15	0
48	39	31	1	-148.1756	2	12	38	0	0	0	0	0
49	25	37	1	-137.1268	1	36	14	0	0	0	0	0
50	49	41	1	-174.8294	1	42	3	0	0	0	5	0

Table C.11: Output of MLE and Bootstrapping For Species 6

#	X	Y	Z	Likelihood	Tree	Bootstrapping Trees						
						1	2	3	4	5	6	7
1	91	92	88	-262.4587	6	0	10	0	0	6	30	4
2	83	89	98	-260.7403	2	16	20	0	0	4	10	0
3	92	90	96	-265.0892	6	6	21	0	3	1	16	3
4	89	94	86	-261.5443	6	2	2	0	0	21	25	0
5	88	92	92	-263.0227	6	8	6	3	0	1	20	12
6	85	93	90	-261.1289	6	18	1	0	3	0	20	8
7	81	88	90	-255.1912	6	10	5	0	1	3	29	2
8	99	93	85	-263.7675	6	0	2	7	6	1	32	2
9	97	94	94	-267.8767	6	18	1	2	0	11	18	0
10	91	86	99	-262.896	6	16	14	0	0	3	17	0
11	92	85	1	-257.457	1	44	5	0	0	0	1	0
12	94	81	97	-260.6156	2	6	22	3	0	0	19	0
13	91	88	84	-257.3399	6	1	2	0	0	0	45	2
14	89	88	83	-255.4895	6	1	2	0	0	0	41	6
15	85	88	1	-257.4199	1	46	1	0	0	0	3	0
16	93	82	92	-257.953	6	15	6	0	1	0	28	0
17	95	83	94	-261.8645	2	6	28	1	0	0	15	0
18	99	90	93	-266.1267	6	9	11	6	0	3	19	2
19	89	89	89	-260.2993	6	8	5	0	0	0	31	6
20	84	99	94	-264.7136	7	1	0	0	21	7	11	10
21	88	92	89	-261.8238	6	6	4	0	0	10	27	3
22	100	81	99	-262.2264	2	15	19	0	0	0	16	0
23	100	89	81	-259.439	6	0	3	0	2	4	41	0
24	92	91	91	-263.7023	6	10	3	2	0	5	26	4
25	90	90	97	-265.1253	1	17	17	0	1	10	5	0
26	89	90	79	-254.4658	6	0	0	0	0	21	29	0
27	84	76	92	-250.8328	2	3	30	0	0	0	17	0
28	84	93	80	-260.9466	2	3	43	0	0	0	4	0
29	83	81	72	-252.545	1	30	1	0	0	7	12	0
30	84	89	97	-261.7969	1	23	6	0	0	16	5	0
31	88	92	95	-263.9593	6	5	20	2	0	0	6	17
32	92	100	74	-265.7532	1	46	0	1	0	3	0	0
33	86	87	89	-256.9973	6	6	6	0	0	4	34	0
34	86	92	91	-261.2618	6	0	14	0	3	1	23	9
35	94	95	86	-264.2491	6	1	2	1	1	20	25	0
36	93	98	86	-265.1457	6	2	3	16	0	2	21	6
37	93	92	89	-263.8931	6	3	5	0	0	0	35	7
38	77	93	65	-255.5657	4	0	0	1	38	2	1	8
39	98	94	87	-265.2253	6	1	2	0	0	24	22	1
40	100	96	88	-267.0711	6	5	2	4	13	2	19	5
41	93	80	99	-259.646	1	14	19	0	0	0	16	1
42	98	91	87	-263.909	6	0	5	0	2	15	28	0
43	93	95	91	-266.379	6	8	5	0	0	14	18	5
44	88	79	91	-255.8811	2	7	33	0	0	0	10	0
45	98	86	85	-258.8824	6	1	3	0	0	2	43	1
46	83	97	85	-259.6556	6	3	0	1	0	8	20	18
47	93	89	83	-258.1886	6	3	0	0	0	5	41	1
48	87	93	88	-260.8696	6	0	6	3	0	2	29	10
49	92	94	84	-261.965	6	6	0	3	0	3	32	6
50	96	80	82	-251.3317	6	1	10	0	0	0	36	3

Table C.13: Output of MLE and Bootstrapping For Species 7

						Bootstrapping Trees						
#	X	Y	Z	Likelihood	Tree	1	2	3	4	5	6	7
1	34	59	16	-101.7346	6	12	1	0	1	0	31	5
2	24	12	1	-91.8256	1	21	3	1	0	0	25	0
3	35	9	1	-100.3913	1	15	6	0	0	0	29	0
4	38	21	1	-131.3423	2	19	30	0	0	0	1	0
5	34	13	1	-108.8242	1	19	26	0	0	0	5	0
6	33	17	1	-115.355	2	24	24	0	0	0	2	0
7	46	9	95	-113.1156	6	38	0	0	0	0	12	0
8	33	13	1	-107.2158	2	13	27	0	0	0	10	0
9	37	10	1	-106.3809	2	6	36	0	0	0	8	0
10	32	9	1	-96.42451	1	26	0	0	0	0	24	0
11	40	66	18	-115.6861	6	7	11	1	0	0	24	7
12	39	5	1	-95.28758	1	7	4	0	0	0	39	0
13	41	22	1	-137.3243	2	13	37	0	0	0	0	0
14	29	17	1	-109.4137	2	1	46	0	1	0	2	0
15	42	14	1	-121.5165	2	8	30	0	0	0	12	0
16	44	15	1	-125.6722	2	12	32	0	0	0	6	0
17	36	13	1	-110.7425	2	22	26	0	0	0	2	0
18	51	18	1	-139.9728	1	28	21	0	0	0	1	0
19	36	24	1	-134.9198	1	24	23	2	0	0	1	0
20	37	86	16	-115.2225	6	9	1	11	0	0	22	7
21	47	74	14	-120.5699	6	8	0	0	0	24	18	0
22	33	59	25	-112.2048	6	11	9	0	0	7	23	0
23	46	9	95	-113.1156	6	28	7	0	0	0	15	0
24	28	17	1	-107.8379	2	12	37	0	0	0	1	0
25	31	9	1	-96.06089	2	10	20	0	0	0	20	0
26	45	15	1	-127.2751	1	39	6	0	0	0	5	0
27	42	12	1	-117.5299	1	21	26	0	0	0	3	0
28	40	7	90	-99.86903	6	16	10	0	0	0	24	0
29	39	10	1	-109.2252	1	28	1	0	0	0	21	0
30	38	11	1	-110.2676	1	20	18	0	0	0	12	0
31	38	16	1	-121.4926	2	0	48	0	0	0	2	0
32	41	14	1	-119.8187	1	32	5	0	0	0	13	0
33	32	18	1	-116.3441	1	33	12	0	0	0	5	0
34	33	17	1	-115.355	2	18	30	0	0	0	1	1
35	33	54	27	-112.0044	6	14	4	1	0	0	28	3
36	40	14	1	-118.0355	2	0	49	0	0	0	1	0
37	40	15	1	-121.9622	2	18	29	0	0	0	3	0
38	33	14	1	-109.3905	1	48	0	0	0	0	2	0
39	41	9	97	-107.3101	6	10	17	0	0	0	23	0
40	48	10	96	-118.8854	6	22	12	0	0	0	16	0
41	43	11	1	-116.122	2	13	29	0	0	0	8	0
42	42	14	1	-120.8174	1	37	7	0	0	0	6	0
43	31	18	1	-115.9558	1	38	11	0	0	0	1	0
44	36	10	1	-105.5934	1	29	0	0	0	0	21	0
45	39	9	1	-105.5675	1	30	1	0	0	0	19	0
46	45	20	1	-136.0699	2	0	50	0	0	0	0	0
47	44	13	1	-123.499	1	30	7	0	0	0	13	0
48	33	8	97	-94.62442	6	6	23	0	0	0	21	0
49	49	17	1	-135.9212	1	50	0	0	0	0	0	0
50	48	16	1	-132.1047	2	13	36	0	0	0	1	0

Table C.15: Output of MLE and Bootstrapping For Species 8

#	X	Y	Z	Likelihood	Tree	Bootstrapping Trees						
						1	2	3	4	5	6	7
1	92	43	1	-215.9176	5	0	0	0	0	48	2	0
2	39	92	96	-210.8875	7	0	0	11	14	0	14	11
3	94	86	46	-216.7043	6	0	0	2	0	2	45	1
4	88	93	36	-203.1026	6	0	0	12	1	1	25	11
5	89	95	33	-198.9998	6	0	0	6	7	5	22	10
6	82	84	36	-194.9145	6	0	0	0	0	6	43	1
7	87	40	91	-212.4041	5	0	0	0	3	26	19	2
8	30	98	70	-194.8125	3	0	0	34	7	5	3	1
9	96	83	37	-201.1066	6	0	0	0	0	9	41	0
10	82	79	44	-202.8893	6	0	0	0	1	3	46	0
11	100	68	47	-204.9632	6	0	0	0	1	0	49	0
12	92	94	39	-211.157	6	0	0	16	6	0	22	6
13	88	84	27	-181.3157	6	0	0	0	7	2	39	2
14	98	81	43	-210.0034	6	0	0	4	1	1	43	1
15	93	50	1	-226.3554	5	0	0	0	0	43	7	0
16	100	89	39	-209.4273	6	0	0	0	4	15	31	0
17	98	92	42	-216.3164	6	0	0	6	2	14	27	1
18	86	75	36	-189.8272	6	2	0	0	0	1	46	1
19	94	82	33	-193.497	6	0	0	0	1	7	41	1
20	88	87	43	-211.2417	6	0	0	0	2	0	45	3
21	98	85	46	-217.1179	6	0	0	0	1	8	41	0
22	84	59	56	-199.3025	6	0	1	0	0	0	49	0
23	84	86	27	-180.9526	6	0	0	5	0	0	41	4
24	88	82	33	-191.5627	6	0	0	2	0	3	44	1
25	78	84	32	-185.8248	6	0	1	0	0	16	28	5
26	93	99	37	-209.6579	6	0	0	3	13	10	21	3
27	82	79	48	-209.1725	6	0	0	0	0	11	39	0
28	98	91	48	-224.1896	6	0	0	9	7	1	32	1
29	100	34	80	-204.9301	5	0	0	1	2	36	11	0
30	87	88	39	-204.3293	6	0	0	4	0	8	38	0
31	84	91	37	-201.2653	6	0	0	0	0	15	30	5
32	85	75	50	-209.2676	6	0	0	0	0	3	46	1
33	85	94	40	-209.2057	6	0	0	15	0	5	18	12
34	79	34	97	-198.1706	5	0	0	0	5	17	17	11
35	94	40	85	-213.7044	5	0	0	7	0	23	20	0
36	85	94	48	-221.7773	6	0	0	8	0	4	31	7
37	92	76	50	-213.78	6	0	0	0	0	0	50	0
38	96	93	35	-205.014	6	0	0	1	12	3	33	1
39	92	47	94	-226.0431	5	0	0	5	1	18	25	1
40	100	80	39	-203.3114	6	0	0	3	3	0	44	0
41	98	88	38	-206.7672	6	0	0	8	1	2	37	2
42	84	35	92	-202.2462	5	0	0	0	1	21	26	2
43	85	32	1	-194.4408	5	0	0	0	1	33	16	0
44	86	83	46	-211.5321	6	0	0	1	1	0	40	8
45	79	91	41	-205.5043	6	0	0	3	1	1	30	15
46	97	92	43	-218.1619	6	0	0	9	1	4	33	3
47	99	85	37	-204.0754	6	0	0	2	1	4	42	1
48	99	83	49	-220.6521	6	0	0	0	7	0	42	1
49	90	83	43	-209.0192	6	0	0	2	0	0	45	3
50	89	76	44	-205.0404	6	0	0	0	0	1	49	0

Table C.17: Output of MLE and Bootstrapping For Species 9

#	X	Y	Z	Likelihood	Tree	Bootstrapping Trees						
						1	2	3	4	5	6	7
1	100	33	1	-198.8	2	11	39	0	0	0	0	0
2	88	58	59	-203.7293	6	1	4	0	0	1	44	0
3	92	42	76	-198.4817	6	10	5	0	0	0	35	0
4	81	40	1	-203.8534	1	31	17	0	0	0	2	0
5	73	37	1	-193.3293	2	12	38	0	0	0	0	0
6	88	41	84	-200.767	6	21	0	0	0	0	29	0
7	99	39	1	-208.9679	1	40	1	0	0	0	9	0
8	85	29	1	-187.1267	1	44	6	0	0	0	0	0
9	87	30	1	-189.3122	1	44	6	0	0	0	0	0
10	100	66	43	-197.9175	6	1	2	0	0	0	47	0
11	88	26	1	-182.4548	1	49	1	0	0	0	0	0
12	99	44	88	-211.8012	6	8	15	0	0	0	27	0
13	100	35	1	-201.6319	2	9	41	0	0	0	0	0
14	84	34	1	-196.4414	2	9	30	0	0	0	11	0
15	88	32	83	-194.6365	2	18	11	0	0	0	21	0
16	82	24	1	-177.0125	1	32	18	0	0	0	0	0
17	97	41	74	-197.3487	6	1	19	0	0	0	30	0
18	88	32	86	-188.0126	6	17	6	0	0	0	27	0
19	85	32	1	-192.0535	2	1	32	0	0	0	17	0
20	100	33	1	-198.8	1	39	11	0	0	0	0	0
21	96	66	37	-188.0565	6	0	0	0	0	0	50	0
22	93	29	1	-190.4122	2	3	47	0	0	0	0	0
23	93	31	1	-193.9686	2	2	34	0	0	0	14	0
24	92	41	80	-200.0261	6	20	1	0	0	0	29	0
25	87	29	1	-189.3821	1	27	23	0	0	0	0	0
26	75	36	1	-194.6329	2	12	27	0	0	0	11	0
27	90	42	1	-209.7861	2	28	22	0	0	0	0	0
28	100	38	83	-209.1723	1	18	13	0	0	0	19	0
29	82	31	1	-189.424	1	45	5	0	0	0	0	0
30	86	29	1	-187.139	2	12	38	0	0	0	0	0
31	91	42	76	-198.2958	6	7	10	0	0	0	33	0
32	84	35	71	-181.436	6	11	11	0	0	0	28	0
33	90	36	1	-202.4038	1	31	0	0	0	0	19	0
34	85	31	1	-190.3905	1	27	23	0	0	0	0	0
35	97	30	1	-194.3694	1	23	3	0	0	0	24	0
36	92	29	1	-190.4406	2	2	24	0	0	0	24	0
37	92	31	91	-190.5578	6	13	7	0	0	0	30	0
38	99	61	67	-219.1667	6	6	0	0	0	0	44	0
39	90	41	81	-199.6342	6	22	1	0	0	0	27	0
40	92	30	98	-193.8498	1	19	13	0	0	0	18	0
41	100	33	1	-200.6317	1	22	18	0	0	0	10	0
42	91	40	67	-209.6254	2	17	22	0	0	0	11	0
43	87	33	1	-193.9264	2	3	47	0	0	0	0	0
44	80	34	1	-193.6801	1	31	7	0	0	0	12	0
45	86	40	1	-205.4719	2	1	46	0	0	0	3	0
46	84	36	66	-179.2351	6	4	17	0	0	0	29	0
47	89	37	1	-203.529	2	7	33	0	0	0	10	0
48	94	35	1	-201.7131	2	18	27	0	0	0	5	0
49	83	32	1	-191.6586	2	5	29	0	0	0	16	0
50	78	29	1	-184.7142	2	21	29	0	0	0	0	0

Table C.19: Output of MLE and Bootstrapping For Species 10

						Bootstrapping Trees						
#	X	Y	Z	Likelihood	Tree	1	2	3	4	5	6	7
1	44	85	76	-202.884	6	0	4	0	0	1	39	6
2	42	85	93	-211.2028	2	8	30	0	0	1	8	3
3	34	82	62	-195.3521	2	11	26	0	0	0	10	3
4	37	79	1	-197.9703	1	30	15	0	0	0	2	3
5	43	85	91	-210.9492	6	16	9	0	0	2	19	4
6	43	84	82	-214.8921	5	1	2	0	0	36	8	3
7	38	95	78	-202.9054	6	7	0	0	0	13	24	6
8	39	78	90	-201.422	2	14	26	0	0	0	9	1
9	35	84	91	-199.7616	1	25	3	0	0	8	10	4
10	27	80	95	-182.3007	1	21	4	0	0	12	13	0
11	44	89	88	-219.0326	5	8	1	0	0	30	4	7
12	43	90	96	-217.5369	1	24	10	0	0	1	9	6
13	43	94	84	-213.9991	6	4	4	0	0	1	22	19
14	34	97	1	-201.194	5	0	26	0	0	24	0	0
15	38	82	94	-201.5769	6	6	21	0	0	0	18	5
16	33	84	98	-199.9368	5	2	5	0	0	21	8	14
17	59	95	87	-237.2268	6	9	3	0	0	7	18	13
18	34	88	79	-200.2015	1	28	7	0	0	13	2	0
19	36	87	92	-200.9975	6	13	5	0	0	7	18	7
20	91	93	40	-210.9511	7	12	1	0	0	0	3	34
21	38	96	86	-206.9125	6	15	0	0	0	3	18	14
22	75	96	43	-209.3823	7	0	0	0	0	10	11	29
23	84	93	40	-208.1239	7	0	12	0	0	3	15	20
24	43	86	85	-215.9105	5	7	15	0	0	22	6	0
25	41	90	86	-209.0247	6	10	1	0	0	0	17	22
26	84	88	25	-177.5517	7	0	1	0	0	13	5	31
27	42	90	86	-214.8097	2	7	30	0	0	0	5	8
28	43	76	92	-203.8467	6	6	10	0	0	0	32	2
29	39	98	59	-211.1535	1	29	0	0	0	18	1	2
30	43	77	1	-206.6592	2	0	32	0	0	0	16	2
31	37	83	81	-193.7525	6	11	0	0	0	7	31	1
32	48	91	90	-221.4293	6	7	9	0	0	2	20	12
33	39	91	81	-202.8691	6	0	9	0	0	3	27	11
34	36	89	96	-203.7121	2	7	17	0	0	0	7	19
35	37	89	90	-205.6328	2	4	31	0	0	7	6	2
36	45	95	80	-219.8896	2	2	34	0	0	9	1	4
37	32	98	85	-197.5059	6	3	8	0	0	8	13	18
38	37	96	94	-209.0634	6	5	12	0	0	6	12	15
39	33	78	73	-191.2261	1	37	6	0	0	2	5	0
40	34	87	1	-196.429	2	1	44	0	0	0	1	4
41	28	92	80	-182.9268	6	1	3	0	0	11	28	7
42	39	84	91	-203.6565	6	7	9	0	0	7	27	0
43	42	90	93	-213.6786	6	18	4	0	0	14	11	3
44	38	94	94	-208.613	6	6	21	0	0	2	11	10
45	29	75	1	-181.514	1	27	21	0	0	0	2	0
46	44	79	96	-211.0616	2	12	19	0	0	3	14	2
47	36	80	88	-198.2488	1	22	7	0	0	0	11	10
48	34	77	97	-194.0765	1	11	10	0	0	4	24	1
49	40	91	87	-207.797	6	2	17	0	0	8	16	7
50	37	93	92	-209.358	5	8	14	0	0	24	2	2

Table C.21: Output of MLE and Bootstrapping For Species 11

#	X	Y	Z	Likelihood	Tree	Bootstrapping Trees						
						1	2	3	4	5	6	7
1	90	89	1	-257.3126	5	0	0	0	0	50	0	0
2	88	95	73	-264.3213	5	0	0	0	0	49	0	1
3	100	84	92	-265.5595	5	0	0	3	15	28	3	1
4	89	89	77	-263.0444	5	1	0	0	0	46	2	1
5	77	84	87	-254.4649	5	0	1	0	2	32	0	15
6	86	92	96	-264.8572	5	11	1	0	0	17	7	14
7	95	85	81	-264.6294	5	0	0	6	2	35	0	7
8	86	88	1	-257.8523	5	0	0	0	0	48	0	2
9	93	85	1	-257.5943	5	0	0	0	0	47	2	1
10	99	84	67	-261.861	5	0	0	1	14	34	0	1
11	93	92	82	-259.9979	6	0	0	0	0	18	32	0
12	100	90	89	-266.8634	2	0	24	2	0	1	23	0
13	92	92	1	-263.0604	5	0	0	1	0	47	1	1
14	90	92	80	-264.0771	5	0	0	2	0	39	0	9
15	92	83	86	-262.3847	5	0	0	0	0	44	6	0
16	86	82	94	-260.612	5	1	1	3	0	21	14	10
17	83	90	79	-260.4577	5	0	1	0	0	41	0	8
18	93	93	95	-268.1476	5	0	2	1	16	24	1	6
19	100	93	90	-268.7422	5	0	1	6	2	37	4	0
20	81	94	84	-261.0329	5	0	0	1	0	35	1	13
21	96	90	93	-267.7426	5	2	0	11	3	26	0	8
22	99	94	84	-264.0676	6	0	2	0	8	18	22	0
23	94	94	1	-264.2636	5	0	0	1	0	48	0	1
24	92	94	90	-267.4719	5	0	7	1	0	36	3	3
25	79	85	90	-256.9485	5	1	0	0	0	40	7	2
26	85	86	97	-262.2198	5	5	0	0	1	23	15	6
27	87	100	90	-265.9814	3	4	0	29	1	2	14	0
28	99	98	95	-269.8984	7	8	3	6	4	18	3	8
29	94	94	92	-267.8835	5	5	2	3	1	33	1	5
30	82	93	86	-256.8358	6	0	6	0	0	9	24	11
31	91	71	84	-252.4726	5	0	0	8	1	35	1	5
32	85	98	85	-263.9753	5	9	0	0	0	37	0	4
33	82	93	96	-261.6059	7	0	0	5	14	15	2	14
34	80	98	84	-257.5334	6	6	0	1	1	9	18	15
35	85	94	92	-264.9388	5	4	14	0	0	25	6	1
36	96	95	96	-268.5283	6	13	1	0	12	5	15	4
37	85	79	1	-253.7218	5	0	0	1	0	46	0	3
38	86	97	98	-266.5605	5	8	2	3	0	19	5	13
39	96	81	86	-261.6255	5	0	0	3	1	40	4	2
40	91	91	77	-263.4343	5	1	3	0	0	40	6	0
41	89	88	1	-259.0789	5	1	0	0	0	47	2	0
42	95	100	92	-269.6541	5	1	12	3	1	32	0	1
43	95	100	92	-269.6541	5	1	10	8	0	30	0	1
44	83	86	95	-261.3153	5	0	3	0	0	25	12	10
45	91	94	94	-266.0418	6	3	9	1	0	15	20	2
46	79	97	87	-261.0684	5	6	1	0	0	39	4	0
47	97	100	71	-266.6365	2	0	40	5	0	4	1	0
48	95	98	96	-269.7616	5	21	0	3	2	20	4	0
49	99	90	72	-265.1235	5	0	0	0	7	43	0	0
50	92	85	73	-523.8821	5	0	0	0	7	43	0	0

Table C.23: Output of MLE and Bootstrapping For Species 12

#	X	Y	Z	Likelihood	Tree	Bootstrapping Trees						
						1	2	3	4	5	6	7
1	38	74	47	-156.9636	6	0	5	0	0	2	42	1
2	34	61	59	-151.5548	6	8	0	0	0	3	39	0
3	39	59	1	-183.8123	4	0	0	0	27	0	1	22
4	26	59	1	-162.5323	2	12	34	0	0	0	4	0
5	51	44	1	-183.1418	1	27	16	0	0	0	7	0
6	100	39	1	-210.2916	1	35	4	0	0	0	11	0
7	63	33	1	-179.906	3	0	0	44	6	0	0	0
8	2	34	1	-75.85782	3	0	0	25	0	0	25	0
9	10	23	1	-85.02215	3	0	2	26	18	0	4	0
10	100	87	1	-256.6182	3	0	0	50	0	0	0	0
11	93	78	58	-254.2328	2	17	31	0	0	0	2	0
12	33	9	1	-98.43591	1	19	5	0	0	0	26	0
13	37	80	85	-193.9695	6	11	2	0	0	0	32	5
14	20	27	1	-114.0821	5	0	0	4	11	30	0	5
15	33	11	93	-102.1064	6	12	3	0	0	0	35	0
16	100	91	13	-156.0322	6	2	0	1	23	5	19	0
17	31	37	1	-145.4116	2	8	42	0	0	0	0	0
18	12	1	1	-33.42918	6	0	0	0	0	0	50	0
19	40	48	1	-173.5823	1	28	4	0	0	0	18	0
20	64	25	1	-165.5601	2	20	30	0	0	0	0	0
21	69	42	1	-198.2024	3	0	0	33	13	0	0	4
22	15	3	93	-50.72198	6	0	0	0	0	0	48	2
23	58	63	63	-218.7085	5	0	0	0	0	36	4	10
24	36	81	33	-142.0009	7	0	1	6	2	9	5	27
25	71	78	79	-240.7922	3	0	0	31	13	0	0	6
26	59	35	1	-182.0754	5	0	0	0	0	47	0	3
27	18	99	73	-162.1283	6	0	6	0	0	12	14	18
28	39	25	1	-138.5634	4	0	0	4	46	0	0	0
29	23	61	1	-158.1598	4	0	0	8	38	0	1	3
30	21	90	1	-173.8831	1	35	1	0	0	14	0	0
31	91	64	1	-239.0003	1	37	12	0	0	0	1	0
32	65	65	1	-221.2489	2	11	34	0	0	0	5	0
33	84	80	37	-194.6155	7	1	1	0	0	0	4	44
34	96	72	69	-232.3161	7	1	0	1	0	0	0	48
35	24	77	68	-155.5866	7	0	0	11	0	0	8	31
36	18	23	1	-102.6277	2	27	21	0	0	0	0	2
37	81	93	22	-171.9866	6	0	0	0	11	2	31	6
38	73	92	17	-155.4487	6	1	2	7	0	5	22	13
39	77	65	11	-134.5695	7	4	0	2	1	0	3	40
40	34	52	1	-168.0538	1	46	0	0	0	0	4	0
41	68	74	86	-229.9397	7	0	0	9	8	0	0	33
42	88	15	1	-158.6905	3	0	0	44	4	0	0	2
43	20	35	1	-125.5121	5	1	2	2	0	34	9	2
44	54	48	1	-195.5823	5	0	0	0	0	37	9	4
45	51	3	1	-99.16743	5	1	0	0	0	33	16	0
46	7	2	73	-29.24904	6	1	0	0	0	0	49	0
47	23	71	1	-166.2416	2	10	37	0	0	0	2	1
48	3	66	1	-111.1896	5	16	0	0	0	28	1	5
49	15	20	1	-89.51904	4	0	0	2	41	0	6	1
50	29	17	1	-109.4154	2	3	47	0	0	0	0	0

Table C.25: Output of MLE and Bootstrapping For Species 13

#	X	Y	Z	Likelihood	Tree	Bootstrapping Trees						
						1	2	3	4	5	6	7
1	96	96	1	-260.6364	5	0	0	0	0	50	0	0
2	87	87	1	-256.5815	5	0	0	0	0	50	0	0
3	90	91	96	-264.9974	6	0	24	0	0	17	8	1
4	96	93	1	-259.5159	5	0	0	0	0	50	0	0
5	76	94	80	-257.1479	5	2	17	0	0	28	3	0
6	94	94	86	-267.5805	5	0	5	4	0	38	1	2
7	94	94	86	-267.5805	5	0	3	1	0	40	2	4
8	86	91	1	-259.3127	5	0	0	0	0	49	0	1
9	86	91	1	-259.3127	5	0	0	0	0	49	0	1
10	98	87	1	-258.4493	5	0	0	0	0	50	0	0
11	97	86	58	-262.0381	5	2	0	0	0	41	7	0
12	81	96	1	-255.6457	5	0	1	0	0	49	0	0
13	89	96	1	-260.9528	5	0	0	0	0	50	0	0
14	97	92	74	-265.9751	5	0	0	0	4	46	0	0
15	89	93	98	-266.0555	7	0	0	16	12	12	0	10
16	96	84	83	-263.5615	5	0	0	4	7	37	0	2
17	84	90	75	-260.7656	5	1	0	0	2	45	0	2
18	91	93	52	-263.3753	5	0	0	2	0	46	0	2
19	89	88	94	-265.0576	5	0	1	1	0	36	6	6
20	89	95	1	-260.1911	5	3	0	0	0	46	1	0
21	90	98	92	-267.5807	1	19	0	0	1	21	8	1
22	86	100	94	-267.0338	5	22	12	0	0	15	1	0
23	91	89	1	-252.7296	5	0	0	0	0	50	0	0
24	91	95	84	-266.1805	5	0	1	0	3	44	1	1
25	91	87	81	-262.1567	5	0	0	0	2	43	3	2
26	94	89	94	-266.9669	5	10	0	0	1	22	14	3
27	87	90	1	-260.1674	5	0	0	0	0	47	3	0
28	99	94	1	-264.7834	5	0	0	1	0	47	2	0
29	96	94	1	-259.1229	5	1	0	0	0	49	0	0
30	88	93	75	-264.0387	5	0	2	0	0	46	2	0
31	88	97	41	-263.577	5	0	1	0	0	49	0	0
32	100	87	71	-263.5017	5	0	0	8	0	41	1	0
33	93	90	1	-261.1094	5	0	0	1	2	45	1	1
34	100	90	70	-264.6104	5	0	0	1	3	44	2	0
35	88	87	66	-260.9968	5	0	3	0	0	43	4	0
36	89	90	88	-263.9141	5	5	0	0	0	33	12	0
37	89	92	1	-258.9139	5	0	0	0	0	50	0	0
38	90	82	53	-257.3827	5	0	0	0	0	49	1	0
39	89	95	1	-262.786	5	3	0	0	0	47	0	0
40	79	100	70	-259.6643	5	3	0	0	0	38	1	8
41	96	96	87	-267.9845	5	0	3	21	0	25	0	1
42	94	92	75	-265.0796	5	2	0	0	0	45	3	0
43	93	84	1	-256.7057	5	0	0	0	0	50	0	0
44	95	78	1	-256.0822	5	0	0	0	0	46	4	0
45	91	88	1	-260.6567	5	0	3	0	0	44	2	1
46	96	97	1	-264.277	5	1	0	3	0	46	0	0
47	95	89	1	-260.8843	5	0	0	3	2	41	0	4
48	78	91	85	-258.8997	5	0	0	0	0	36	0	14
49	96	89	79	-264.9376	5	0	0	0	18	32	0	0
50	82	100	88	-263.7315	5	4	8	0	0	30	2	6

Table C.27: Output of MLE and Bootstrapping For Species 14

#	X	Y	Z	Likelihood	Tree	Bootstrapping Trees						
						1	2	3	4	5	6	7
1	31	95	92	-197.6258	3	0	0	18	19	7	3	3
2	99	41	77	-215.9791	5	0	0	14	4	32	0	0
3	100	32	86	-200.3309	5	0	0	2	17	29	2	0
4	33	96	1	-198.921	4	0	0	0	46	2	1	1
5	43	96	91	-217.2957	7	0	0	5	7	3	16	19
6	81	35	90	-201.0396	5	0	0	9	0	27	8	6
7	81	35	90	-201.0396	5	0	0	8	0	30	10	2
8	90	33	92	-201.6932	5	0	0	0	1	28	21	0
9	90	33	92	-201.6932	5	0	0	0	1	29	15	5
10	35	86	65	-199.2041	4	0	0	1	33	9	0	7
11	78	47	85	-218.3054	5	0	0	6	0	28	3	13
12	39	92	91	-209.5341	7	0	0	1	17	6	9	17
13	41	98	91	-215.525	7	0	0	0	23	14	10	3
14	95	33	1	-198.2749	5	0	0	0	1	45	3	1
15	81	93	46	-215.6896	6	0	0	1	0	12	36	1
16	100	40	76	-213.8155	5	0	0	0	8	29	13	0
17	77	96	37	-201.4007	6	0	0	1	2	2	22	23
18	98	88	34	-199.7107	6	0	0	0	1	14	35	0
19	28	89	91	-186.6237	7	0	0	23	0	2	18	7
20	89	92	37	-205.4709	6	0	0	0	0	22	27	1
21	97	91	34	-202.3893	6	0	0	7	3	7	31	2
22	87	90	42	-211.0816	6	0	0	2	0	0	33	15
23	94	27	82	-189.5293	5	0	0	6	4	29	11	0
24	93	91	32	-196.995	6	0	0	0	5	10	31	4
25	42	98	89	-218.5081	3	0	0	21	7	19	1	2
26	45	83	95	-213.3051	7	0	0	19	6	6	1	18
27	40	96	92	-213.156	7	0	0	4	14	12	10	10
28	90	34	1	-200.7073	5	0	0	0	0	39	11	0
29	81	34	87	-198.6061	5	0	0	2	0	25	23	0
30	28	97	96	-192.5115	7	0	0	15	4	13	15	3
31	31	93	76	-196.6857	4	0	0	5	24	11	4	6
32	36	91	92	-202.7408	7	0	0	11	15	10	2	12
33	86	88	40	-206.2666	6	0	0	0	0	21	29	0
34	34	93	1	-198.0731	3	0	0	40	4	6	0	0
35	91	94	30	-194.3314	6	0	0	0	3	17	30	0
36	96	36	1	-204.7749	5	0	0	2	5	43	0	0
37	30	99	88	-197.8503	4	0	0	1	22	14	11	2
38	84	48	57	-222.1471	5	0	0	1	2	38	3	6
39	96	34	88	-205.0373	5	0	0	6	11	27	5	1
40	89	32	75	-198.5875	5	0	0	0	1	32	15	2
41	90	45	89	-221.3378	5	0	0	0	2	28	18	2
42	80	41	64	-208.459	5	0	0	0	8	27	1	14
43	97	91	41	-213.8568	6	0	0	20	0	2	27	1
44	84	39	87	-208.9545	5	0	0	1	3	34	1	11
45	96	99	40	-216.6821	6	0	0	15	5	9	21	0
46	40	93	84	-207.2174	7	0	0	1	8	0	17	24
47	41	96	96	-216.249	7	0	0	10	8	21	5	6
48	38	81	70	-201.4985	4	0	0	13	32	0	1	4
49	82	33	1	-191.5509	5	0	0	0	0	49	1	0
50	89	42	67	-214.2176	5	0	0	0	2	41	6	1

C.2

Table C.29: Bootstrapping For Species 2 with DNA Sequences

#	X	Y	Z	Likelihood	MLE Tree	Bootstrapping Trees						
						1	2	3	4	5	6	7
1	81	96	83	-261.6993	2	1	45	0	0	3	1	0
2	96	87	77	-262.8739	1	37	1	0	0	3	9	0
3	94	90	83	-264.9898	1	29	0	2	0	7	12	0
4	100	92	98	-268.6477	2	16	14	9	9	0	2	0
5	89	94	86	-264.5168	1	36	6	1	0	1	4	2
6	95	97	1	-262.589	1	47	0	0	1	0	2	0
7	89	95	84	-264.5598	1	38	3	0	0	1	7	1
8	100	94	1	-262.4725	1	50	0	0	0	0	0	0
9	94	86	94	-263.298	1	27	16	0	0	1	5	1
10	88	81	1	-252.5431	1	47	0	0	0	0	3	0

Table C.31: Bootstrapping For Species 8 with DNA Sequences

#	X	Y	Z	Likelihood	MLE Tree	Bootstrapping Trees						
						1	2	3	4	5	6	7
1	84	95	41	-211.1605	6	0	0	2	0	15	29	4
2	90	86	39	-205.0766	6	0	0	0	0	26	24	0
3	78	43	97	-213.0961	5	0	0	3	0	21	11	15
4	100	81	47	-216.3957	6	0	0	0	3	2	45	0
5	88	84	43	-208.6663	6	0	0	0	2	4	44	0
6	36	86	1	-199.1348	3	0	0	39	0	3	4	4
7	94	91	40	-211.0931	6	0	0	0	0	13	35	2
8	34	100	59	-202.164	4	0	0	1	36	0	13	0
9	90	95	37	-206.5552	6	0	0	1	8	10	26	5
10	100	87	38	-206.8139	6	0	0	4	2	0	43	1

Table C.33: Output of MLE and Bootstrapping For Species 13

#	X	Y	Z	Likelihood	Tree	Bootstrapping Trees						
						1	2	3	4	5	6	7
1	94	91	87	-267.0443	5	0	0	2	2	42	0	4
2	99	91	90	-267.9877	5	1	0	0	20	28	1	0
3	89	100	73	-265.1202	5	0	7	0	0	43	0	0
4	95	92	79	-265.2235	5	2	0	0	0	46	2	0
5	96	84	77	-262.2094	5	0	0	1	2	46	1	0
6	96	93	79	-266.2966	5	2	1	1	0	40	5	1
7	96	91	52	-264.4116	5	1	0	2	0	47	0	0
8	94	94	81	-265.5535	5	5	0	0	2	37	6	0
9	93	99	1	-263.3402	5	0	3	0	0	45	2	0
10	91	99	1	-257.7463	5	0	0	0	0	50	0	0

References

Aeschbacher, S., Beaumont, M. A., and Futschik, A. (2012). A novel approach for choosing summary statistics in approximate bayesian computation. *Genetics*, 192(3):1027–1047.

Åkerborg, Ö., Sennblad, B., Arvestad, L., and Lagergren, J. (2009). Simultaneous bayesian gene tree reconstruction and reconciliation analysis. *Proceedings of the National Academy of Sciences*, 106(14):5714–5719.

Alfaro, M. E., Zoller, S., and Lutzoni, F. (2003). Bayes or bootstrap? a simulation study comparing the performance of bayesian markov chain monte carlo sampling and bootstrapping in assessing phylogenetic confidence. *Molecular Biology and Evolution*, 20(2):255–266.

Allman, E. S., Degnan, J. H., and Rhodes, J. A. (2011a). Determining species tree topologies from clade probabilities under the coalescent. *Journal of theoretical biology*, 289:96–106.

Allman, E. S., Degnan, J. H., and Rhodes, J. A. (2011b). Identifying the rooted species tree from the distribution of unrooted gene trees under the coalescent. *Journal of mathematical biology*, 62(6):833–862.

Allman, E. S., Degnan, J. H., and Rhodes, J. A. (2016). Species tree inference from gene splits by unrooted star methods. *arXiv preprint arXiv:1604.05364*.

Beaumont, M. A. et al. (2010). Approximate bayesian computation in evolution and ecology. *Annual Review of Ecology, Evolution and Systematics*, 41(379-406):1.

Beaumont, M. A., Zhang, W., and Balding, D. J. (2002). Approximate bayesian computation in population genetics. *Genetics*, 162(4):2025–2035.

Bertorelle, G., Benazzo, A., and Mona, S. (2010). Abc as a flexible framework to estimate demography over space and time: some cons, many pros. *Molecular ecology*, 19(13):2609–2625.

Blum, M. G. and Tran, V. C. (2010). Hiv with contact tracing: a case study in approximate bayesian computation. *Biostatistics*, 11(4):644–660.

Bouckaert, R., Heled, J., Kühnert, D., Vaughan, T., Wu, C.-H., Xie, D., Suchard, M. A., Rambaut, A., and Drummond, A. J. (2014). Beast 2: a software platform for bayesian evolutionary analysis. *PLoS Comput Biol*, 10(4):e1003537.

Boykin, L. M., Kubatko, L. S., and Lowrey, T. K. (2010). Comparison of methods for rooting phylogenetic trees: A case study using orcuttieae (poaceae: Chloridoideae). *Molecular phylogenetics and evolution*, 54(3):687–700.

Buzbas, E. O. (2012). On the article titled ?estimating species trees using approximate bayesian computation?(fan and kubatko, molecular phylogenetics and evolution 59: 354–363). *Molecular phylogenetics and evolution*, 65(3):1014–1016.

Buzbas, E. O. and Rosenberg, N. A. (2015). Aabc: Approximate approximate bayesian computation for inference in po pulation-genetic models. *Theoretical population biology*, 99:31–42.

Casella, G. and Berger, R. L. (2002). *Statistical inference*, volume 2. Duxbury Pacific Grove, CA.

Chen, G. K., Marjoram, P., and Wall, J. D. (2009). Fast and flexible simulation of dna sequence data. *Genome research*, 19(1):136–142.

Chifman, J. and Kubatko, L. (2014). Quartet inference from snp data under the coalescent model. *Bioinformatics*, 30(23):3317–3324.

Csilléry, K., Blum, M. G., Gaggiotti, O. E., and François, O. (2010). Approximate bayesian computation (abc) in practice. *Trends in ecology & evolution*, 25(7):410–418.

Degnan, J. H. (2013). Anomalous unrooted gene trees. *Systematic biology*, page syt023.

Degnan, J. H. and Rhodes, J. A. (2015). There are no caterpillars in a wicked forest. *Theoretical Population Biology*, 105:17–23.

Degnan, J. H. and Rosenberg, N. A. (2006). Discordance of species trees with their most likely gene trees. *PLoS Genet*, 2(5):e68.

Degnan, J. H. and Rosenberg, N. A. (2009). Gene tree discordance, phylogenetic inference and the multispecies coalescent. *Trends in ecology & evolution*, 24(6):332–340.

Degnan, J. H. and Salter, L. A. (2005). Gene tree distributions under the coalescent process. *Evolution*, 59(1):24–37.

Drovandi, C. C. and Pettitt, A. N. (2011). Estimation of parameters for macroparasite population evolution using approximate bayesian computation. *Biometrics*, 67(1):225–233.

Drovandi, C. C. and Pettitt, A. N. (2012). Likelihood-free inference for transmission rates of nosocomial pathogens. *Case Studies in Bayesian Statistical Modelling and Analysis*, pages 374–387.

Drummond, A. J., Ho, S. Y., Phillips, M. J., and Rambaut, A. (2006). Relaxed phylogenetics and dating with confidence. *PLoS Biol*, 4(5):e88.

Eck, R. V. and Dayhoff, M. O. (1966). {Atlas of Protein Sequence and Structure}.

Edwards, A. (1964). F. and cavalli-sforza. *LL:'Reconstruction of Evolutionary Trees'.[See Ref. 10, 67-76]*.

Edwards, A. W. (1970). Estimation of the branch points of a branching diffusion process. *Journal of the Royal Statistical Society. Series B (Methodological)*, pages 155–174.

Fan, H. H. and Kubatko, L. S. (2011). Estimating species trees using approximate Bayesian computation. *Molecular phylogenetics and evolution*, 59(2):354–363.

Farris, J. S. (1973). On comparing the shapes of taxonomic trees. *Systematic Biology*, 22(1):50–54.

Farris, J. S. (1977). Phylogenetic analysis under dollo's law. *Systematic Biology*, 26(1):77–88.

Fearnhead, P. and Prangle, D. (2012). Constructing summary statistics for approximate bayesian computation: semi-automatic approximate bayesian computation. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 74(3):419–474.

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

Felsenstein, J. (2001). Taking variation of evolutionary rates between sites into account in inferring phylogenies. *Journal of Molecular Evolution*, 53(4-5):447–455.

Felsenstein, J. (2004). *Inferring phylogenies*, volume 2. Sinauer Associates Sunderland.

Fu, Y.-X. and Li, W.-H. (1997). Estimating the age of the common ancestor of a sample of dna sequences. *Molecular biology and evolution*, 14(2):195–199.

Fukami-Kobayashi, K. and Tateno, Y. (1991). Robustness of maximum likelihood tree estimation against different patterns of base substitutions. *Journal of molecular evolution*, 32(1):79–91.

Goldman, N. (1990). Maximum likelihood inference of phylogenetic trees, with special reference to a poisson process model of dna substitution and to parsimony analyses. *Systematic Biology*, 39(4):345–361.

Gomberg, D. (1966). Bayesian" postdiction in an evolution process. *Unpublished manuscript*.

Guindon, S. and Gascuel, O. (2003). A simple, fast, and accurate algorithm to estimate large phylogenies by maximum likelihood. *Systematic biology*, 52(5):696–704.

Harper, C. W. (1979). A bayesian probability view of phylogenetic systematics. *Systematic Biology*, 28(4):547–553.

Hasegawa, M., Kishino, H., and Saitou, N. (1991). On the maximum likelihood method in molecular phylogenetics. *Journal of molecular evolution*, 32(5):443–445.

Heled, J. and Drummond, A. J. (2010). Bayesian inference of species trees from multilocus data. *Molecular biology and evolution*, 27(3):570–580.

Hornik, K. and Meyer, D. (2009). Generalized and customizable sets in r. *Journal of Statistical Software*, 31(2):1–27.

Huang, H., He, Q., Kubatko, L. S., and Knowles, L. L. (2010). Sources of error inherent in species-tree estimation: impact of mutational and coalescent effects on accuracy and implications for choosing among different methods. *Systematic biology*, 59(5):573–583.

Hudson, R. R. (1983). Properties of a neutral allele model with intragenic recombination. *Theoretical population biology*, 23(2):183–201.

Huelsenbeck, J. P. (1995). Performance of phylogenetic methods in simulation. *Systematic biology*, 44(1):17–48.

Huelsenbeck, J. P., Bollback, J. P., and Levine, A. M. (2002). Inferring the root of a phylogenetic tree. *Systematic biology*, 51(1):32–43.

Huelsenbeck, J. P. and Hillis, D. M. (1993). Success of phylogenetic methods in the four-taxon case. *Systematic Biology*, 42(3):247–264.

Huelsenbeck, J. P. and Kirkpatrick, M. (1996). Do phylogenetic methods produce trees with biased shapes? *Evolution*, pages 1418–1424.

Ihaka, R. and Gentleman, R. (1996). R: a language for data analysis and graphics. *Journal of computational and graphical statistics*, 5(3):299–314.

Jensen, J. D., Thornton, K. R., and Andolfatto, P. (2008). An approximate bayesian estimator suggests strong, recurrent selective sweeps in drosophila. *PLoS Genet*, 4(9):e1000198.

Kashyap, R. and Subas, S. (1974). Statistical estimation of parameters in a phylogenetic tree using a dynamic model of the substitutional process. *Journal of Theoretical Biology*, 47(1):75–101.

Kingman, J. F. C. (1982). The coalescent. *Stochastic processes and their applications*, 13(3):235–248.

Knowles, L. L. and Kubatko, L. S. (2011). *Estimating species trees: practical and theoretical aspects*. John Wiley and Sons.

Kubatko, L. S. and Fan, H. H. (2013). Letter to the editor on the article entitled" Estimating species trees\ using Approximate Bayesian Computation"(Fan

and Kubatko, Mol. Phylogenetics Ev\ ol. 59, 354-363). *Molecular phylogenetics and evolution*, 66(1):438.

Kuhner, M. K. and Felsenstein, J. (1994). A simulation comparison of phylogeny algorithms under equal and unequal evolutionary rates. *Molecular Biology and Evolution*, 11(3):459–468.

Larget, B. R., Kotha, S. K., Dewey, C. N., and Ané, C. (2010). Bucky: gene tree/species tree reconciliation with bayesian concordance analysis. *Bioinformatics*, 26(22):2910–2911.

Laval, G. and Excoffier, L. (2004). Simcoal 2.0: a program to simulate genomic diversity over large recombining regions in a subdivided population with a complex history. *Bioinformatics*, 20(15):2485–2487.

Liu, L. and Pearl, D. K. (2007). Species trees from gene trees: reconstructing bayesian posterior distr ibutions of a species phylogeny using estimated gene tree distributions. *Systematic Biology*, 56(3):504–514.

Liu, L. and Yu, L. (2011). Estimating species trees from unrooted gene trees. *Systematic biology*, 60(5):661–667.

Liu, L., Yu, L., Kalavacharla, V., and Liu, Z. (2011). A bayesian model for gene family evolution. *BMC bioinformatics*, 12(1):426.

Liu, L., Yu, L., Pearl, D. K., and Edwards, S. V. (2009). Estimating species phylogenies using coalescence times among sequences. *Systematic Biology*, 58(5):468–477.

Maddison, W. P. (1997). Gene trees in species trees. *Systematic biology*, 46(3):523–536.

Marjoram, P. and Tavaré, S. (2006). Modern computational approaches for analysing molecular genetic variation data. *Nature Reviews Genetics*, 7(10):759–770.

Mirarab, S., Reaz, R., Bayzid, M. S., Zimmermann, T., Swenson, M. S., and Warnow, T. (2014). Astral: genome-scale coalescent-based species tree estimation. *Bioinformatics*, 30(17):i541–i548.

Nei, M. (1987). *Molecular evolutionary genetics*. Columbia university press.

Neyman, J. (1971). *Molecular studies of evolution: a source of novel statistical problems*.

Nichols, R. (2001). Gene trees and species trees are not the same. *Trends in Ecology & Evolution*, 16(7):358–364.

Nunes, M. A., Balding, D. J., et al. (2010). On optimal selection of summary statistics for approximate bayesian computation. *Statistical applications in genetics and molecular biology*, 9(1):34.

Pamilo, P. and Nei, M. (1988). Relationships between gene trees and species trees. *Molecular biology and evolution*, 5(5):568–583.

Paradis, E., Claude, J., and Strimmer, K. (2004). Ape: analyses of phylogenetics and evolution in r language. *Bioinformatics*, 20(2):289–290.

Pritchard, J. K., Seielstad, M. T., Perez-Lezaun, A., and Feldman, M. W. (1999). Population growth of human y chromosomes: a study of y chromosome microsatellites. *Molecular biology and evolution*, 16(12):1791–1798.

Rambaut, A. and Grassly, N. C. (1997). Seq-gen: An application for the Monte Carlo simulation of DNA sequence evolution along phylogenetic trees. *Comp. Appl. Biosci.*, 13:235–238.

Rannala, B. and Yang, Z. (1996). Probability distribution of molecular evolutionary trees: a new method of phylogenetic inference. *Journal of molecular evolution*, 43(3):304–311.

Rannala, B. and Yang, Z. (2003). Bayes estimation of species divergence times and ancestral population sizes using dna sequences from multiple loci. *Genetics*, 164(4):1645–1656.

Ranwez, V. and Gascuel, O. (2002). Improvement of distance-based phylogenetic methods by a local maximum likelihood approach using triplets. *Molecular Biology and Evolution*, 19(11):1952–1963.

Rasmussen, M. D. and Kellis, M. (2012). Unified modeling of gene duplication, loss, and coalescence using a locus tree. *Genome Research*, 22(4):755–765.

Robert, C. P. (2016). Approximate bayesian computation: A survey on recent results. In *Monte Carlo and Quasi-Monte Carlo Methods*, pages 185–205. Springer.

Roch, S. and Warnow, T. (2015). On the robustness to gene tree estimation error (or lack thereof) of coalescent-based species tree methods. *Systematic biology*, page syv016.

Rogers, J. S. (1997). On the consistency of maximum likelihood estimation of phylogenetic trees from nucleotide sequences. *Systematic biology*, 46(2):354–357.

Rokas, A., Williams, B. L., King, N., and Carroll, S. B. (2003). Genome-scale approaches to resolving incongruence in molecular phylogenies. *Nature*, 425(6960):798–804.

Ronquist, F., Teslenko, M., van der Mark, P., Ayres, D. L., Darling, A., Höhna, S., Larget, B., Liu, L., Suchard, M. A., and Huelsenbeck, J. P. (2012). Mrbayes 3.2: efficient bayesian phylogenetic inference and model choice across a large model space. *Systematic biology*, 61(3):539–542.

Rosenberg, M. S. and Kumar, S. (2001). Traditional phylogenetic reconstruction methods reconstruct shallow and deep evolutionary relationships equally well. *Molecular Biology and Evolution*, 18(9):1823–1827.

Rosenberg, N. A. (2002). The probability of topological concordance of gene trees and species trees. *Theoretical population biology*, 61(2):225–247.

Rosenberg, N. A. (2013). Discordance of species trees with their most likely gene trees: a unifying principle. *Molecular biology and evolution*, page mst160.

Salichos, L. and Rokas, A. (2013). Inferring ancient divergences requires genes with strong phylogenetic signals. *Nature*, 497(7449):327–331.

Salichos, L., Stamatakis, A., and Rokas, A. (2014). Novel information theory-based measures for quantifying incongruence among phylogenetic trees. *Molecular Biology and Evolution*, page msu061.

Semple, C. and Steel, M. A. (2003). *Phylogenetics*, volume 24. Oxford University Press on Demand.

Shoemaker, J. S., Painter, I. S., and Weir, B. S. (1999). Bayesian statistics in genetics: a guide for the uninitiated. *Trends in Genetics*, 15(9):354–358.

Simmons, M. P. and Gatesy, J. (2015). Coalescence vs. concatenation: Sophisticated analyses vs. first principles applied to rooting the angiosperms. *Molecular phylogenetics and evolution*, 91:98–122.

Smouse, P. E. and Li, W.-H. (1987). Likelihood analysis of mitochondrial restriction-cleavage patterns for the human-chimpanzee-gorilla trichotomy. *Evolution*, pages 1162–1176.

Stadler, T. (2014). Treesim: Simulating trees under the birth-death model. r package version 2.0.

Steel, M. (2012). Root location in random trees: A polarity property of all sampling consistent phylogenetic models except one. *Molecular phylogenetics and evolution*, 65(1):345–348.

Sues, H. D. (2016). Amniotes, diversification of. *Encyclopedia of Evolutionary Biology, ed RM Kliman (Oxford: Academic Press, pp.56-62.)*.

Tavaré, S., Balding, D. J., Griffiths, R. C., and Doñnelly, P. (1997). Inferring coalescence times from dna sequence data. *Genetics*, 145(2):505–518.

Than, C., Ruths, D., and Nakhleh, L. (2008). Phylonet: a software package for analyzing and reconstructing reticulate evolutionary relationships. *BMC bioinformatics*, 9(1):1.

Wakeley, J. (2009). *Coalescent theory: an introduction*. Number 575: 519.2 WAK.

Wang, Y. and Degnan, J. H. (2011). Performance of matrix representation with parsimony for inferring species from gene trees. *Statistical Applications in Genetics and Molecular Biology*, 10(1).

Weiss, G. and von Haeseler, A. (1998). Inference of population history using a likelihood approach. *Genetics*, 149(3):1539–1546.

Wheeler, W. C. (1991). Congruence among data sets: A bayesian approach. *Phylogenetic analysis of DNA sequences (MM Miyamoto and J. Cracraft, eds.). Oxford Univ. Press, Oxford, UK*, pages 334–346.

Wickett, N. J., Mirarab, S., Nguyen, N., Warnow, T., Carpenter, E., Matasci, N., Ayyampalayam, S., Barker, M. S., Burleigh, J. G., Gitzendanner, M. A., et al. (2014). Phylotranscriptomic analysis of the origin and early diversification of land plants. *Proceedings of the National Academy of Sciences*, 111(45):E4859–E4868.

Woodhams, M. D., Lockhart, P. J., and Holland, B. R. (2016). Simulating and summarizing sources of gene tree incongruence. *Genome Biology and Evolution*, page evw065.

Wright, S. (1931). Evolution in mendelian populations. *Genetics*, 16(2):97–159.

Wu, Y. (2012). Coalescent-based species tree inference from gene tree topologies under incomplete lineage sorting by maximum likelihood. *Evolution*, 66(3):763–775.

Xi, Z., Liu, L., Rest, J. S., and Davis, C. C. (2014). Coalescent versus concatenation methods and the placement of amborella as sister to water lilies. *Systematic biology*, 63(6):919–932.

Yang, Z. (1994). Statistical properties of the maximum likelihood method of phylogenetic estimation and comparison with distance matrix methods. *Systematic biology*, 43(3):329–342.

Yu, Y., Barnett, R. M., and Nakhleh, L. (2013). Parsimonious inference of hybridization in the presence of incomplete lineage sorting. *Systematic biology*, page syt037.

Yu, Y., Than, C., Degnan, J. H., and Nakhleh, L. (2011). Coalescent histories on phylogenetic networks and detection of hybridization despite incomplete lineage sorting. *Systematic Biology*, 60(2):138–149.

Zhu, S., Degnan, J. H., Goldstien, S. J., and Eldon, B. (2015). Hybrid-lambda: simulation of multiple merger and kingman gene genealogies in species networks and species trees. *BMC bioinformatics*, 16(1):292.

Zhu, S. J. (2012). Hybrid coal.