RECONSTRUCTING POSTERIOR DISTRIBUTIONS OF A SPECIES PHYLOGENY USING ESTIMATED GENE TREE DISTRIBUTIONS

DISSERTATION

Presented in Partial Fulfillment of the Requirements for

the Degree Doctor of Philosophy in the

Graduate School of The Ohio State University

By

Liang Liu, M.S.

* * * * *

The Ohio State University

2006

Dissertation Committee:

Dennis K. Pearl, Adviser

Joseph S. Verducci

Steven N. MacEachern

Daniel Janies

Approved by

Adviser

GRADUATE PROGRAM
in BIOSTATISTICS

© Copyright by

Liang Liu

2006

ABSTRACT

The desire to infer the evolutionary history of a group of species should be more viable now that a considerable amount of multilocus molecular data is available. However, the current molecular phylogenetic paradigm still reconstructs gene trees to represent the species tree. Further, commonly used methods to combine data, such as the concatenation method, the consensus tree method, or the gene tree parsimony method may be biased. In this dissertation, I propose a Bayesian hierarchical model to estimate the phylogeny of a group of species using multiple estimated gene tree distributions such as those that arise in a Bayesian analysis of DNA sequence data. The model employs substitution models used in traditional phylogenetics, but also uses coalescent theory to explain genealogical signals from species trees to gene trees and from gene trees to sequence data, thereby forming a complete stochastic model to simultaneously estimate gene trees, species trees, ancestral population sizes, and species divergence times. The proposed model is founded on the assumption that gene trees, even of unlinked loci, are correlated due to being derived from a single species tree and therefore should be estimated jointly. The method is applied to three multilocus data sets of DNA sequences. The estimates of the species tree topology and divergence times appear to be robust to the prior of the population size, whereas the estimates of effective population sizes are sensitive to the prior used in the analysis. These analyses also suggest that the model is superior to the concatenation method in

fitting these data sets and thus provides a more realistic assessment of the variability in the distribution of species trees that may have produced the molecular information at hand. Future improvements of our model and algorithm should include consideration of other factors that can cause discordance of gene trees and species trees, such as horizontal transfer or gene duplication.

This is dedicated to the people I love, my son, my wife and my parents.

ACKNOWLEDGMENTS

I would never have been able to finish this work without help, support, and encouragement from my wife, Lili Yu. First of all, I would like to thank my advisor Dennis K. Pearl, for his inspiring and encouraging way to guide me to a deeper understanding of knowledge work, and his invaluable comments during the whole work with this dissertation. I will also give special thanks to Scott Edwards for a fruitful collaboration and constructional commends on the model formulation and extension. I am very grateful to Steven MacEachern for his instruction on the Bayesian model selection and Markov Chain Monte Carlo method. I am also very grateful to everyone that have read parts of the manuscript, especially Joseph S. Verducci, Daniel Janies and Tina Henkin. Thanks also to Mathematical Bioscience Institute at The Ohio State University for providing a super computer account. At last, I would like to thank Professor Anthony J. Tosi, Antonis Rokas, Bryan Jennings and Scott Edwards for providing the molecular data.

VITA

1995 ..... B.S. Clinical Medicine

2000 ..... M.S. Neuroscience

2005 ..... M.S. Statistics

PUBLICATIONS

Research Publications

Liang Liu and Guowei Lu. Protective effect of protein-free supernatant of brain homogenate taken from hypoxia preconditioned mice on synaptosome membrane exposed to hypoxia. *Chinese Journal of Neuroscience*, 17(4):373–375, 2001.

Andrea D. Wolfe, C.P. Randle, L. Liu and K.E. Steiner. Phylogeny and biogeography of orobanchaceae. *Folia Geobotanica*, 40(2-3):115, 2005.

FIELDS OF STUDY

Major Field: Biostatistics

Studies in:

Phylogenetics	Prof. Dennis K. Pearl
Bayesian Analysis	Prof. Steven N. MacEachern

TABLE OF CONTENTS

Page

LIST OF TABLES

LIST OF FIGURES

CHAPTER 1

INTRODUCTION

The evolutionary process produces a branching pattern of species. The common ancestors of different species split into two lineages through speciation. The two lineages continue to evolve and split independently except through the uncommon exchange of genetic information via migration or hybridization, etc. By studying inherited species' characteristics such as nucleotides, proteins and other historical evidence, we can reconstruct the evolutionary history of species and represent it as a species tree.

Cladistics is one of the most commonly used techniques to build a species tree [24] [6]. It is based on the fact that individuals of a group share a common evolutionary history, and are more closely related to individuals of the same group than to the individuals in other groups. These groups are identified by sharing unique features which are not present in other species [48]. These shared characteristics are called synapomorphies. A cladistics analysis starts with coding the characters and determining the state of the characters for each taxon [72]. The states of characters for each taxon are then concatenated into a sequence. For each site that represents the state of a particular character along the sequence, a tree is constructed according

to the shared derived characters of taxa. The trees of different sites are often incongruent. The estimate of the species tree is the one that can minimize the number of conflicts that arise from the trees of different sites [113] [80]. Before the invention of advanced molecular biology technologies, morphological and physiological features of species were commonly used as the characters. However, due to the difficulty of measuring morphological or physiological similarity among species, reconstruction of species trees has been highly controversial among biologists.

As the technology of molecular biology appeared and advanced rapidly, tremendous amounts of genetic data became available. It was recognized that building a phylogenetic tree from genetic data was much easier and sometimes more appropriate than other traditional approaches. A variety of evolutionary models arose to explain the mechanism of mutation and other genetic processes [77]. The statistical techniques such as the maximum likelihood [31] [30] and Bayesian inference [129] [82] [73] were adapted to build species trees. Because the datasets generally consisted of the sequences from only one gene, those methods were referred by most people as gene tree reconstruction methods. The gene trees reconstructed by these methods were treated as the estimates of species trees by assuming that the gene trees are identical to the species trees. However, the phylogenetic tree reconstructed from a particular gene may not agree with the species tree. The gene tree reconstruction methods are thus generally not directly applicable to the species tree reconstruction.

As the huge amount of multilocus genetic data has accumulated in the recent 15 years, building a species tree has become a more attainable goal and has been increasingly studied in evolutionary genetics. The concatenation method [55] and consensus tree method [16] are two commonly used techniques that use multilocus

genetic data to estimate species trees. For the consensus tree method, gene trees are inferred separately for each gene, and the consensus tree of these gene phylogenies is used as the estimate of the species tree. On the other hand, the concatenation method concatenates the gene sequences into a super-gene alignment, which is then analyzed to estimate the species tree.

If a dataset consists of equal sizes of molecular characters and morphological characters, the method of consensus appears to imply an equal weighting of molecular data and morphological data [14]. However, the equal weightings implied by the consensus tree method may be difficult to defend [5]. Similarly, for the concatenation method, the species tree estimate is more likely to be determined by the genes with long sequences than by the short genes. The estimate of the species tree is then biased if the gene trees for the long genes happen to have wrong topologies.

In this research, we propose a Bayesian hierarchical model to estimate the phylogeny of a group of closely related species from multilocus molecular data. Our model employs the substitution model and coalescent theory to explain the evolutionary process from species trees to gene trees and from gene trees to sequence data. This thereby forms a statistical approach to estimate gene trees, species trees, ancestral population sizes and divergence times simultaneously.

Before we lay out the Bayesian hierarchical model, we will first make a general introduction to the gene tree reconstruction methods, coalescent theory and species trees estimation methods. We will then discuss the Bayesian hierarchical model formulation and its properties in chapter 2. The Markov Chain Monte Carlo method will be addressed as well in chapter 2. Results from the simulation study will be presented and discussed in chapter 3. In chapter 4, we will apply the new method to

real data. Finally, we conclude with a discussion of the strength and weakness of the proposal and of possible fruitful areas of future research.

1.1 Gene Tree Reconstruction Methods

Genealogical phylogeny, as the words suggest, is the phylogeny of a particular gene from a group of species. When the molecular data such as DNA sequences or protein sequences were first used to infer phylogenies, the amount of data was limited either by the number of species or the number of genes. Most data for a particular study were collected from only one gene. The phylogenetic trees estimated from the single-gene data are in fact gene trees, but they are instead used as the estimate of the species tree by assuming that gene trees and the species trees are identical even though the assumption is not always true. However, gene tree estimation plays a central role in the current species tree reconstruction techniques. It is worth illustrating how to estimate a gene tree before we jump into the topic of the species tree reconstruction method.

The molecular data used for the rest of this chapter are DNA sequences. Readers should be aware that the gene tree reconstruction methods also work for other types of data such as protein sequences, morphological data, behavioral data, and mixture data.

1.1.1 Parsimony Methods

Parsimony methods were among the first methods for inferring gene trees. The general idea of the parsimony method was first introduced in Edwards and Cavalli-Sforza's paper [22]. For the following decades, a huge amount of literature were published to explore the properties of the methods as well as to improve the efficiency

of the algorithm. [39] [35] [27] [80] [100].We will focus on the following aspects: the general idea, algorithm, and the statistical properties of the methods.

The idea of the parsimony methods is quite straightforward: the most parsimonious tree (MP tree) is an evolutionary tree that requires minimum number of evolutionary steps to explain the observed pattern. The gene tree is estimated by the most parsimonious tree. [40]

We have to search all possible trees in order to find the MP tree. The tree space of interest is huge with $\pi_{k=1}^{n}(2k+1)$ topologies, which is already $1.78 * 10^{42}$ for just n=30 species. It is impossible to search over all the possible trees if the number of species is large and search algorithms often become trapped at local optima. The same problem also occurs in the maximum likelihood method.

Searching for the MP tree can be characterized as a minimization problem. Let $f(T)$ be the minimal number of evolutionary changes required for a given tree T and Ω_T be the tree space. Our primary goal is to find a tree T^* such that $f(T^*) = \min\{f(T), T \in \Omega_T\}$. The Newton-Raphson Method is not applicable for this situation because of the discontinuity of $f(T)$.

There are basically three techniques for the heuristic search in the tree space. The starting tree is crucial for any of those three techniques. A good starting point can make the convergence faster and deliver the global optimum. On the contrary, a bad starting point may result in being trapped locally and unable to reach the global optimum.

The first technique adopts the neighborhood joining tree as the staring point. Small rearrangement of branches is made on the starting tree. We call the staring tree the old tree and the tree after the rearrangement the new tree. The new tree is

accepted if its $f(T)$ is smaller than the old tree's $f(T)$. Otherwise, the new tree is rejected and the chain returns to the old tree. The algorithm continues until there is no rearrangement to improve $f(T)$. Common branch rearrangement techniques include nearest-neighbor interchange, sub tree pruning and re-grafting, tree bisection and reconnection [114] [69] [2], sectorial searches, and tree fusing [44].

The second technique chooses a different strategy for the starting point. The characters are re-weighted and the modified data are analyzed to estimate the starting tree [94]. The starting tree and the original data are used together to find the MP tree.

The first two techniques are deterministic in the sense that after the starting point and the tree search strategy are specified, they will always end up with the same MP tree. It does not have the property that a better tree is for sure to be found if the algorithm runs long enough. This motivates the use of stochastic search techniques such as "simulated annealing method" [75] [109] for which a new tree is accepted if it has a smaller $f(T)$ or the new tree is accepted with a small probability if it has a little bit higher value of f(T). This method can facilitate the algorithm to jump out of the local trap. It ensures that the global optimum tree will be found if the chain runs long enough and eventually has searched all the trees in the tree space.

The statistical properties of the parsimony algorithm has been extensively investigated by Farris [28] and Felsenstein [34] [32] who attempted to connect the parsimony method with the maximum likelihood method. They have shown that under regular conditions, the parsimony algorithm minimizing $f(T)$ is equivalent to maximizing a likelihood function. But the conditions they assume are not always satisfied. The parsimony method may be inconsistent for the cases where the rate of change is too

high [56]. However, the parsimony method is a good approximation to the maximum likelihood method if the rate of change is not too high. It is attractive because it is faster and simpler than the maximum likelihood method and it is generally more applicable to non-molecular data where realistic probability models are difficult to construct.

1.1.2 Maximum Likelihood Methods

In 1969, Jukes and Cantor proposed the first stochastic model for the change of nucleotides [59] in which there was only one parameter that was the rate of change at equilibrium. Kimura [62] introduced a two parameter model in which the transition and transversion had two different rates. The Kimura two-parameter model and the Jukes-Cantor model place great restrictions on the DNA sequences. Both models assume equal expected frequencies for all four bases. More widely used models, such as F84 [37] [67] and HKY [49], allow arbitrary base frequencies which is apparently more realistic. These substitution models can be used to calculate pairwise distances among DNA sequences from different species. These distances can then be used to find an optimal tree, such as the gene tree that minimizes the sum of the squared errors of the distances in the tree with the distances calculated from the substitution model. This distance method does not consider the gene tree as a parameter in the likelihood function. In 1981, Felsenstein provided a likelihood function $f(D|T)$ of the DNA sequences given the gene tree [33]. The estimate of the gene tree is obtained by maximizing $f(D|T)$ with respect to T.

In the Felsenstein paper, it was assumed that the sites are independent and that the evolutionary processes along different lineages are independent. These assumptions imply that the likelihood $f(D|T)$ is just the product of probability distribution of each site [33].

$$f(D|T) = \prod_{i=1}^{m} f(D^{(i)}|T)$$

where m is the number of sites and $f(D^{(i)}|T)$ is the likelihood function for site i.

Once we know how to calculate the likelihood, the estimation becomes a standard mathematical problem: maximizing the likelihood function over the entire parameter space. The parameters here include not only gene trees but also the parameters in the substitution models and any parameters used to model the correlation among sites. Now we encounter the same problem as we faced in the parsimony method: how do we search in the tree space? Apparently, all the search techniques mentioned for the parsimony method are also applicable for the current situation.

The maximum likelihood estimates are generally consistent in that the estimates converge to the true gene trees as the length of sequences goes to infinity if the underlying model is correct. The proof of consistency requires two steps. It is first proved that the true tree produces the maximum likelihood if there is infinite number of characters. The second step deals with the uniqueness and continuity of the MLE [123].

Maximum likelihood is a model-based approach. Unfortunately, it is impossible to obtain the true model except in a special case such as simulation for which the true tree is pre-specified. Statistical tests are often employed to find the best model for the data. The likelihood ratio test is one of the most commonly used techniques to

select the model from among of a hierarchy of models that most appropriately fits the data. However, choosing the model with high likelihood value may lead to the one that is unnecessarily complex. In addition, it is inappropriate to use the likelihood ratio test to select the topology of the gene tree because the likelihood ratio test requires the models to be nested. This has led many investigations to consider model selection criterion such as the Akaike information criterion (AIC) [1]. AIC consists of two components, goodness of fit and the complexity of the model.

$$AIC = -2lnL + 2\Lambda$$

where lnL is the log likelihood that measures the goodness of fit of the model, and Λ is the number of parameters, which represents the complexity of the model. AIC does not require the nested structure of the models.

Another commonly used model selection approach, cross validation, is based on minimizing the prediction error. However, cross validation involves intensive computation, which dramatically limits its application to tree building projects.

We have so far discussed the maximum likelihood method for building a single gene tree. If the data have multiple genes from the same set of species, how do we estimate the gene tree for each gene? Should we estimate gene trees separately pretending they are independent or jointly assuming they are correlated? Most current techniques assume genes are independent and estimate gene trees separately.

It is obvious that all genes should more or less share the similar evolutionary history since they all are from the same species. A more appropriate assumption is that the gene trees are conditionally independent given the species tree, but they are marginally dependent. We will see this structure in our Bayesian hierarchical model.

1.1.3 Bayesian Method

The Bayesian method is different from the likelihood method in that it treats parameters as random variables and assumes prior distributions on them. The prior represents the initial guess about the parameters distribution without the data. The initial guess will be improved by the data using Bayes theorem:

$$f(\theta|D) = \frac{f(D|\theta)f(\theta)}{f(D)}$$

Making inference on the parameter θ is based on the posterior probability distribution $f(\theta|D)$ which is the combination of likelihood $f(D|\theta)$ and prior $f(\theta)$. We have already mentioned the likelihood function in the section 1.1.2, but choosing a good prior is still an open problem. If the data carry strong information about the parameters, the prior will not strongly affect the posterior $f(\theta|D)$ and the Bayesian inference is close to the maximum likelihood method. A non-informative prior may be used if there is no information about the prior distribution but it is sometimes difficult to find. If the parameters are correlated, a non-informative prior for one parameter may affect the prior of another parameter such that the prior of the second parameter is not non-informative any more. Additionally, different users may have different priors. As a result, software like phylogenetic program MrBayes provides a variety of priors for users. Users may choose different priors and compare the results to see if they are sensitive to the priors.

Bayesian inference is made upon the posterior distribution. The normalization constant in the posterior probability is intractable in most cases. Numerical methods such as Markov Chain Monte Carlo (MCMC) are then implemented to estimate the

posterior distribution of parameters. Under some regular conditions, MCMC converges to the posterior distribution of interest. However, the convergence rate may be slow as the number of parameters increases and the algorithm may not converge in a limited period of time. It is also possible that MCMC can get trapped around a local optimum. Monitoring convergence is an important and challenging problem for MCMC.

The implementation of our Bayesian hierarchical model involves intensive use of the popular phylogenetic program MrBayes. MrBayes uses MCMC to estimate the posterior probability of gene trees as well as the parameters in the substitution model. MrBayes provides a wide range of substitution models and priors for users. Moreover, it can handle different types of data, nucleotides data, protein data, morphological data or mixture data. MrBayes gives users an option to partition the data set and put different evolutionary models on the partitions. Partitions may be independent or share the same parameters. Users are allowed to specify different priors for different partitions. If users want to reconstruct gene trees for each gene, the gene trees of different genes are assumed independent by MrBayes except in the case that users choose clock:coalescence or clock:birthdeath as the prior of the gene trees and make all gene trees share the same parameter, effective population size θ. By sharing the same parameter, the gene trees for different genes become correlated.

Our Bayesian hierarchical model is able to estimate gene trees and species trees simultaneously. As a result, a number of extra parameters have been introduced into the model. To improve the efficiency of the program, we split the whole process into two consecutive steps, from DNA sequences to gene trees and from gene trees to species trees. For the first step, we have taken advantage of MrBayes to estimate

the posterior distribution of the gene trees given the data. Users are allowed to use any models or priors in MrBayes to estimate the posterior distribution of gene trees. The gene tree output will then be used to estimate species trees by another program "Bayesian estimate of species Tree" (BEST).

To select between two models, M_1 and M_2, we need to calculate the posterior probability of the models respectively and select the model with the highest posterior probability. If the prior for models is uniform, choosing the model with highest posterior probability is equivalent to the Bayes Factor approach in which the marginal probabilities of the data for the two models are calculated to determine the model that most fit the data

$$BF = \frac{f(D|M_1)}{f(D|M_2)}$$

.

where $f(D|M_i)$ is the marginal probability of data given the model M_i. It is trivial to show that if the prior odds is equal to 1, the posterior odds is equal to the BF.

$$\frac{f(M1|D)}{f(M2|D)} = \frac{f(D|M_1)}{f(D|M_2)} \times \frac{f(M1)}{f(M2)} = \frac{f(D|M_1)}{f(D|M_2)} = BF$$

There are at least two approaches to estimate Bayes Factor.

1. Reversible Jump MCMC [11] [10]. The model indicator M is added into the sampling scheme, so that at convergence the MCMC forms a sample from the marginal posterior distribution of M (model index), $f(M|D)$. The BF is estimated by the ratio of the number of M_1 to the number of M_2 in the sample. This method is simple but it suffers from the violations of the convergence condition if the models use different parameterization.

2. Estimate the marginal probability of the data for different models from the output of MCMC. The simplest way to estimate the marginal probability of the data is the harmonic mean method [91]. Let $f(\theta|D)$ be the posterior distribution of θ. The marginal probability distribution

$$f(D) = [\int_{\theta} \frac{1}{f(D|\theta)} f(\theta|D) d\theta]^{-1}$$

which suggests the estimate

$$\hat{f(D)} = [\frac{1}{m} \sum_{i=1}^{m} \frac{1}{f(D|\theta_i)}]^{-1}$$

where θ_i is the i^{th} sample from the posterior distribution of θ given data. The harmonic mean estimate converges almost surely to the true $f(D)$ as $m \to \infty$, but it does not generally satisfy the central limit theorem because its variance is usually infinite. Although the harmonic mean approach is unstable and sensitive to the small likelihood value, it often gives results that are accurate enough for interpretation on the logarithmic scale [9] [107]. Modification of the harmonic mean method has been suggested to get around its instability. Interested readers may find details in the papers of Newton and Raftery (1994) [91] and Meng and Wong (1993) [83].

1.2 Coalescent Theory

The coalescent process has been extensively used to estimate important parameters such as ancestral population sizes, migration rates and divergence times in evolutionary genetics. Its underpinning is a "looking backwards" stochastic process used to infer the retrospective behavior of species. This powerful theory for

studying various quantities pertaining to gene genealogies was initiated by Kingman (1982a) [64] [65] [63], Hudson (1982) [52] and Tajima (1983) [115]. Coalescent theory has provided the conceptual framework for studies of DNA sequence variation within species. It is the source of essential tools for making inferences about mutation, recombination, population structure, and natural selection from DNA sequence data. Fu (1999) [42], in his paper, listed three useful features of coalescent theory.

1. Coalescent theory is a sample-based theory.

2. Coalescent theory can develop highly efficient algorithms for simulating population samples under various population genetics models.

3. Coalescent theory is suitable for molecular genetics data analysis.

Although coalescent theory has been generalized to deal with a variety of types of data, there are still many unsolved problems for future research. The challenging problems include making inference for models with selection or for the models that involve multiple factors such as selection, migration, mutation etc [120].

Definition [64]: Let ϵ_n be the finite set of equivalence relations on $1, 2, ...n$. For $R \in \epsilon_n$, $|R|$ is the number of equivalence classes of R. $\{R_t; t \geq 0\}$ is a continuous time Markov chain with state space ϵ_n and is called an n-coalescent if R_0 is the identity relation $\Delta = \{\{1\}, \{2\}, ...\{n\}\}$ and the transition rate is

$q_{\xi,\eta} = \lim_{h\to 0} \frac{P(R_{t+h}=\eta | R_t=\xi)}{h}$ for $\eta, \xi \in \epsilon_n, \xi \neq \eta$ are given by

$$q_{\xi,\eta} = \begin{cases} 1 & \text{if } \xi \prec \eta; \\ 0 & \text{otherwise.} \end{cases}$$

$\xi \prec \eta$ means that η is obtained from ξ by combining two of the equivalence classes so that $\xi \prec \eta$ impies $|\xi| = |\eta| + 1$.

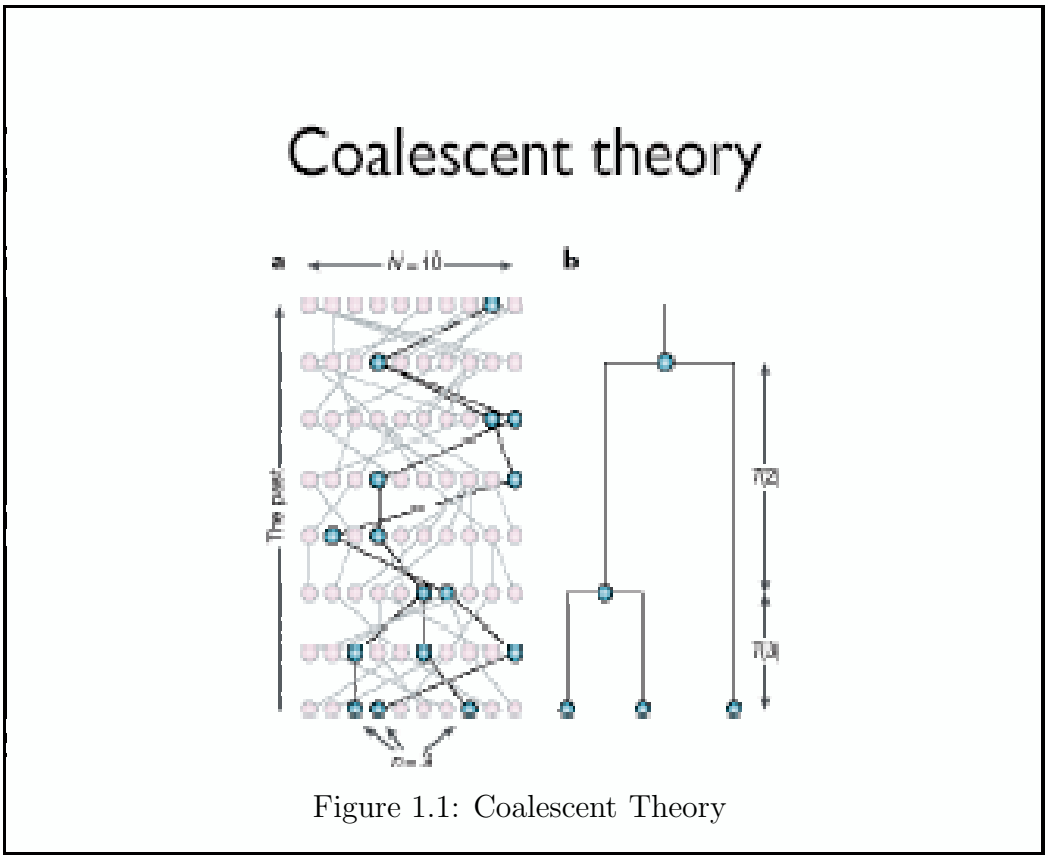

Figure 1.1: Coalescent Theory

To illustrate the idea, consider a sample of N individuals in the population (Figure 1.1). Assume that the population follows the Wright-Fisher model (section 1.2.1). The population size is constant over time, which is 10 in Figure 1.1. For each generation, the individuals may be reproduced and be present in the following generation or may not be reproduced and thereby lost from the population.

If we look backwards in time and start with k ($k = 3$ in Figure 1.1) individuals in the sample at generation 0 (current generation), we see that two individuals have a common ancestor at generation 2 (two generations ago). As we go further back in time, the number of ancestors either decreases by 1 or remains the same at each

preceding generation. Each reduction in the number of ancestors is called a coalescent event. The coalescent process continues until the number of common ancestors reaches 1. The coalescent process can be represented by a tree as that in Figure 1.1.

Consider two individuals at generation 0, A and B. Under the assumptions of the Wright-Fisher model, the individuals at generation 1 are equally likely to be the ancestor of A or B. Thus, the probability that A and B had a common ancestor at generation 1 is $\frac{1}{N}$ and the probability that A and B had a common ancestor at generation $(t+1)$ is then

$$\frac{1}{N}(1-\frac{1}{N})^t \approx \frac{1}{N}e^{-\frac{t}{N}}$$

The exponential approximation is fairly good when N is large. If it is a sample of three individuals, the probability that three individuals have three distinct ancestors at the previous generation is $(1-\frac{1}{N})(1-\frac{2}{N})$. In general, the probability that k individuals have k distinct ancestors at the previous generation is

$$Pr(k) = \prod_{i=1}^{k-1}(1-\frac{i}{N}) \approx 1 - \frac{k!}{(k-2)!2!N}$$

Therefore, the probability that there are k ancestors at generation t while $(k-1)$ ancestors at generation $(t+1)$ is given by

$$Pr(k)^t * (1-Pr(k)) \approx \frac{k!}{(k-2)!2!N}\exp(-\frac{k!}{(k-2)!2!N}t)$$

This approximation is good if $k \ll N$.

Note that here the "individual" does not have to be the "organism" in the population. It may be any "individual" that follows the assumptions of the model. For example, the individuals may be copies of genes. Then the tree generated from the

coalescent process is a gene tree. The number of copies of genes is $2N$ for the diploid population of size N. The probability that there are k ancestors at generation t while $(k-1)$ ancestors at generation $(t+1)$ is given by

$$Pr(k)^t * (1 - Pr(k)) \approx \frac{k!}{(k-2)!4N} \exp(-\frac{k!}{(k-2)!4N}t)$$

Assuming constant population size is not biologically realistic, but in evolutionary genetics the effects of variable population sizes are usually taken into account by computing the effective population size N_e which was introduced by Sewall Wright [125] [126]. It was defined as the number of breeding individuals in an idealized population that would show the same amount of dispersion of allele frequencies under random genetic drift or the same amount of inbreeding as the population under consideration. Effective population size has been found to work surprisingly well [120]. For this reason, we will use N_e instead of N for the rest of the dissertation.

1.2.1 Wright-Fisher model.

The simplest coalescent process is the Wright-Fisher model (Fisher (1922) [38], Wright (1931) [125]) describes the evolution of a two-allele locus in a population of constant size undergoing random mating and ignoring the effects of mutation or selection.

The Wright-Fisher model assumes:

1. Population is of constant size N.

2. Let ν_j be the number of offspring reproduced by the j^{th} member in the population at generation t and assume that the $\{\nu_1, \nu_2..., \nu_N\}$ has a multinomial distribution.

3. Non-overlapping generations.

4. Random mating.

5. No mutation.

6. No selection.

The ancestral process of the Wright-Fisher model has been studied in several papers, including Karlin and McGregor (1972) [61], Cannings (1974) [8], Watterson (1975) [124], Griffiths (1980) [46], Kingman (1980) [63] and Tavare (1984) [117]. Let $\{R_t, t \geq 0\}$ be the ancestor relations at generation t. R_t has two components— the number of ancestors and the structure of the ancestors. D_t is the number of ancestors and R_k is the structure of the ancestors. Note that $D_t = |R_t|$. Let g_{kj}=P(k individuals have j distinct parents). It can be shown that

$g_{k,k-1} = \begin{pmatrix} k \\ 2 \end{pmatrix} + O(N^{-2})$.

$g_{k,k} = 1 - \begin{pmatrix} k \\ 2 \end{pmatrix} + O(N^{-2})$.

$\Rightarrow G_N = I + N^{-1}Q + O(N^{-2})$, where G_N is the transition matrix and the entries in Q are given by $q_{k,k} = -\begin{pmatrix} k \\ 2 \end{pmatrix}, q_{k,k-1} = \begin{pmatrix} k \\ 2 \end{pmatrix}$. $\{D_t\}$ can be approximated by pure death process with transition rate

$\lim_{h\to 0} \frac{P(D_{t+h}=a|D_t=k)}{h} = \begin{cases} \frac{k(k-1)}{2} & \text{if } a = k-1; \\ 0 & \text{if } a \neq k, k-1; \end{cases}$

R_k is called jump process by Kingman(1982a) [64]. Under the Wright-Fisher model, R_k is a Markov chain with transition probability

$P(R_k = \eta | R_k = \xi) = \begin{cases} \frac{2}{k(k-1)} & \text{if } \xi \prec \eta; \\ 0 & \text{otherwise} \end{cases}$.

Kingman [64] proved that $\{D_t\}$ and $\{R_k\}$ are independent. R_t is the combination of D_t and R_k with $R_k = R_{D_t}$. Therefore, R_t has transition rate.

$$q_{\xi\eta} = \begin{cases} 1 & \text{if } \xi \prec \eta; \\ 0 & \text{otherwise.} \end{cases} .$$

and R_0 is the identity relation $\Delta = \{\{1\}, \{2\}, ...\{n\}\}$. $\{R_t\}$ is a coalescent process.

1.2.2 Canning model.

The Canning model assumes that $\sum_{k=1}^{n} \nu_k = N$ and ν_j are exchangable. The Wright-Fisher model is a special case of the Canning model. The transition rate of the coalescent process for the Canning model is

$$g_{kj} = \begin{pmatrix} N \\ k \end{pmatrix}^{-1} \begin{pmatrix} N \\ j \end{pmatrix} \sum_{b \in \Delta_j^k} E \begin{pmatrix} \nu_1 \\ b_1 \end{pmatrix} .... \begin{pmatrix} \nu_j \\ b_j \end{pmatrix}$$

Kingman [65] showed that if $Var(\nu_1) \to \sigma^2$ as $N \to \infty$ and the moments of ν_1 are bounded, then $g_{k,k-1} = \begin{pmatrix} k \\ 2 \end{pmatrix} \sigma^2 N^{-1} + O(N^{-2})$. Consequently, $\{R_t\}$ (genealogy of the sample n) of the Canning model can be well approximated by the coalescent process with the new time scale $t' = \sigma^2 t$.

1.2.3 Variable population size.

The assumption of constant population size is far from the truth. To assess the impact of the fluctuation of the population size over time, let N_x be the population size at generation j. Define the population size function $f_N(x) = \frac{N_j}{N}$, where N is the present population size. We are interested in the behavior of $f_N(x)$ when the N is large. We assume that $\lim_{N\to\infty} f_N(x) = f(x)$, $f(x) > 0$ for all $x \geq 0$. It can be shown that $\{R_t^\nu \, t \geq 0\}$ is still a coalescent process, where ν denotes the variable population size.

Let T_k^2 be the event that two individuals have two distinct ancestors at generation k. It is easy to see that Prob(two individuals have a common ancestor at generation 1) $= \frac{1}{N_1}$, and $P(T_1^2) = 1 - \frac{1}{N_1}$, $P(T_k^2) = \prod_{j=1}^{k} \left(1 - \frac{1}{N_1}\right)$, so $log(P(T_k^2)) = \sum_{j=1}^{k} log\left(1 - \frac{1}{N_j}\right)$, indicating

$$\lim_{N\to\infty} log(P(T_k^2)) = \lim_{N\to\infty} \sum_{j=1}^{k} \frac{1}{N_1} = -\int_0^t \frac{1}{f(x)}\, dx$$

.

Therefore, P(T_t^2)= $exp\,(-\Lambda(t))$, where $\Lambda(t) = \int_0^t \frac{1}{f(x)}\, dx$ and time t is rescaled in units of N. T_t^2 has transition rate $\frac{1}{f(t)}$. Let T_t^i be the event that any two of i individuals have distinct ancestors at generation t, where t is in units of N generations. It can be shown that T_t^i has probability P$(T_t^i) = \begin{pmatrix} i \\ 2 \end{pmatrix} exp\,(-\Lambda(t))$. in which the transition rate is $\begin{pmatrix} i \\ 2 \end{pmatrix} \frac{1}{f(t)}$. It follows that $R^{\nu}(t) = R(\Lambda\,(t))$.

1.2.4 Mutation and coalescent theory.

Due to mutation and natural selection, the DNA sequences of the current species may be quite different from those of their ancestors. The current sequences were evolved from their ancestral sequences along the lineages in a genealogical tree. The polymorphism of the current sequences contains important information about the genealogical phylogeny and may therefore be used to estimate the gene tree. The coalescent process with mutation is one of the models to explain the polymorphisms of the observed molecular sequences.

If there is no selection during evolution, the population reproduction structure may not be affected by mutation. Mutation can be superimposed on the genealogical

tree, which corresponds to the coalescent process with mutation. There are many ways to model mutations along the genealogical tree.

(i) Infinite-many-sites model [124].

The infinitely-many-sites model assumes that

1. Mutation occurs only once at a site.

2. Mutations follow a Poisson process at rate of $\theta/2$ independently along each branch of the gene tree, where $\theta = 2N\mu$.

The Infinite-many-sites model is motivated by so-called segregating sites data. The random variable of interest is the number of segregating sites (S_n). By the assumption of the infinite-many-sites model, the number of segregating sites is equal to the number of mutations (M_n) in the sequence. Therefore, S_n and M_n have the same distribution. Moreover, if the total length of the gene tree (L_n) is given, M_n follows a Poisson distribution with mean $\frac{\theta L_n}{2}$. It can be shown that

$P(S_n = m) = \frac{n-1}{\theta} \sum_{l=1}^{n-1} (-1)^{l-1} \begin{pmatrix} n-2 \\ l-1 \end{pmatrix} \left(\frac{\theta}{\theta+l}\right)^{m+1}$.

The parameter θ can be estimated by maximizing the function $P(S_n = m)$.

(ii) Infinite-many-alleles model.

Assumption:

(1) Each mutation produces a new allele that has not been seen in the population before.

(2) Mutations follow a Poisson process with rate $\theta/2$ independently along each branch, where $\theta = 2N\mu$.

This model is motivated by so-called allozyme frequency data in which only frequencies of genes are observed. The distribution of frequencies was obtained by

Ewens [25], while Kingman [66] derived the same distribution using the coalescent process with mutation.

(iii) Substitution model.

The data of interest are full DNA sequences. The distribution of the genealogical tree is derived from the coalescent theory. This distribution is treated as the prior of gene trees. Given the gene tree, the change of nucleotides along each lineage follows the substitution model, which is the likelihood function. The gene tree is then estimated via Bayesian techniques with the prior from the coalescent theory and the likelihood from the substitution model.

1.2.5 Recombination and coalescent.

Recombination is fairly common for biparental species. If recombination occurs, the current allele has two ancestors—the allele from its mother and the allele from its father, with the number of ancestors increasing by 1. If the current number of ancestors is k, the number of ancestors one generation ago may be k+1, k, or k-1. It is assumed that the recombination rate is ρ. The transition rate

$k \to k+1$ at rate $\frac{k\rho}{2}$ (recombination occurred)

$k \to k-1$ at rate $\frac{k(k-1)}{2}$ (coalescent event occurred)

which is a birth and death process and may be reduced to a pure death process if the recombination rate ρ is zero.

Consider two linked loci, A and B. Recombination data contain the information about the correlation of A and B, which is an important parameter in evolutionary genetics. Griffiths (1997) [47] proposed a so-called ancestor recombination graph (ARG) method in which the ordinary coalescent theory is generalized to include

recombination in the evolutionary process. The likelihood derived from the coalescent theory with recombination demonstrates the probability distribution of the observed molecular data given the genealogical phylogeny. The recombination rate ρ can be estimated by maximizing this likelihood function.

1.2.6 Selection and coalescent.

In the previous sections, we assumed that the alleles were selectively neutral. However, the neutral model is sometimes not adequate to fit real data. In this section, we will discuss the coalescent model with selection. We assume the constant population size and random mating still hold.

The reproduction system of a population may change if natural selection occurs. The individuals with high survival ability tend to have more offspring. The probability that two individuals have a common ancestor is not equal to $\frac{1}{2N}$. Instead, it depends on the mutation rate and the distribution of alleles in the population.

Consider a simple situation in which there are only two alleles A_1 and A_2. The population is divided into two subpopulations by the two types of alleles. One consists of A_1 only and the other consists of A_2 only. The coalescent process within each subpopulation is the regular coalescence without selection because the individuals of each subpopulation have the same ability to produce offspring. However, two coalescent processes have different coalescent rates which depend on the mutation rate and selection parameter. Neuhauser and Krone (1997) [89] have developed a powerful tool known as an ancestral selection graph which enables us to estimate the selection parameter from the data. The coalescent theory with selection was further extended to allow for recombination between neutral site and selected site [60].

These many generalizations of the basic Wright-Fisher model allow for the possibility that some of the simplifying assumptions in the current work may be relaxed in future research.

There have been numerous reviews on the coalescent theory over the past decade, including Hudson (1991, 1992) [53] [54], Ewens (1990) [26], Tavare (1993) [119], Donnelly and Tavare (1995) [19], Fu and Li (1999) [42], Li and Fu (1999) [74]and Neuhauser and Tavare (2001) [90]. Nordborg (2001) [96] has the most comprehensive review of the coalescent theory with selfing, substructure, migration and selection etc.

1.3 Species Tree Estimation

A species tree depicts the pattern of branching of species lineages via the process of speciation. As the ancestral species split, the gene copies within the ancestral population are split into the separated populations of descent. Within each population, gene trees continue branching and descending through time, indicating that gene trees are contained within the branches of the species phylogeny.

As for the multilocus molecular data, gene trees reconstructed for each locus may not agree with the species tree in two respects. First of all, the divergence times of the gene tree are earlier than those of the species tree. Secondly, the topology of a gene tree may be incongruent with that of species tree if deep coalescence, gene duplication and loss or horizontal transfer occurs [45] [3] [99] [116] [127] [20][78] [79].

1. Deep coalescence. The incongruence between species tree and gene tree may be caused by deep coalescence when the gene splitting time was much earlier than the speciation time (Figure 1.2). The probability of deep coalescence is determined by the population size and branch length of the population of

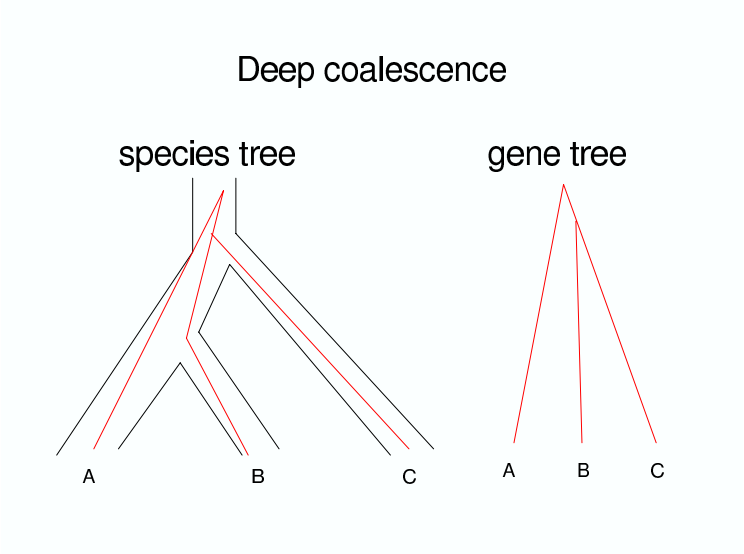

Figure 1.2: Deep Coalescence.

interest in the species tree. The coalescent processes in different populations are assumed to be independent. The probability distribution of the gene tree given the species tree is the product of the likelihood of the independent coalescent processes in the populations.

2. Gene duplication/loss. When gene duplication occurs, each of the duplicated genes evolves along its own evolutionary pathway. The duplicated genes produce a gene family. If the sampled copies in the extant species are paralogous instead of orthologous, it may cause the conflict of the gene tree and the species tree (Figure 1.3).

3. Horizontal gene transfer. When horizontal transfer occurs, a gene from one species introgresses into another species (Figure 1.4). Horizontal transfer may be induced by a virus or mite. The probability of horizontal transfer depends on the

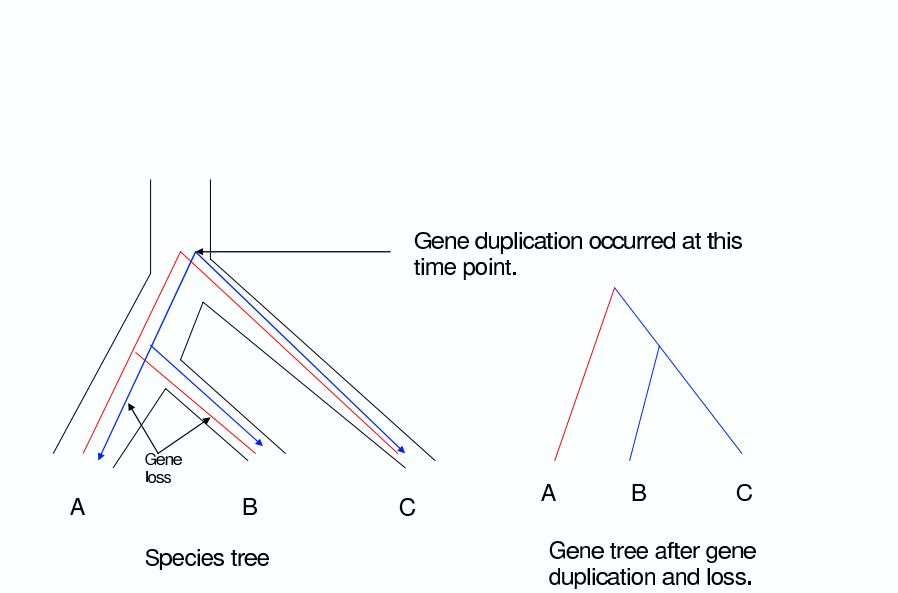

Figure 1.3: Gene duplication and loss.

distance of the original and receiving species. If two species are phylogenetically distant, it is unlikely for them to have horizontal transfer.

There are two levels of errors in reconstructing a species tree from molecular sequences. The first level is in estimating gene trees while the second level is in the process of estimating a species tree from gene trees. Reconstructing the species tree from just one gene tree can yield an erroneous species tree, even if the gene tree is reconstructed correctly. Multiple gene data contain much more information about the species tree than data from a single gene. Using multiple genes may dramatically reduce estimation errors. There are currently three approaches to estimate the species tree using multilocus data.

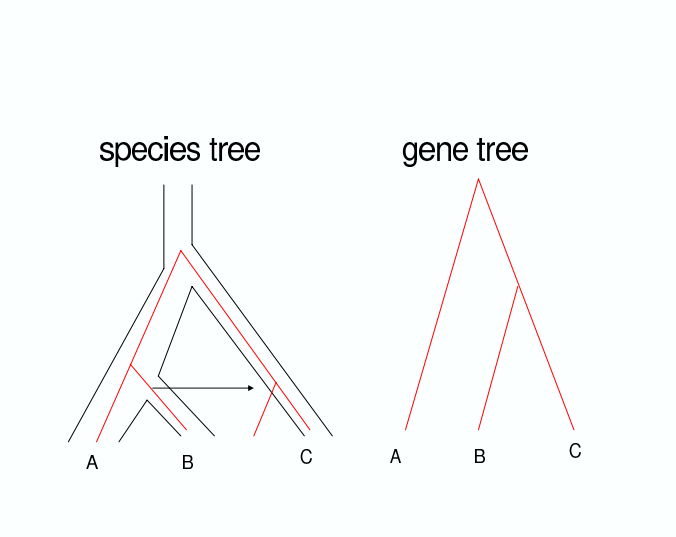

Figure 1.4: Horizontal Transfer. The topology of the species tree is (A,B)C. The gene is transferred from species B to species C, resulting in a gene tree with the topology (B,C)A that is incongruent with the topology of the species tree.

1.3.1 Combined-data approach.

Kluge [68] proposed a method to combine DNA sequences from multiple genes. He argued that all available gene sequences be connected into a single sequence for analysis. The most commonly used combined-data approach is the concatenation method. The concatenated sequences represent all of the unpartitioned evidence at hand [68]. Some of the information contained in the combined sequences may be lost if the data is partitioned and summarized by a consensus tree. The concatenation method has been shown by a simulation study [43] to yield more accurate trees, even when the sequences have evolved with very different substitution patterns.

However, the concatenation approach is not appropriate for genes with different histories. In some cases, the estimate produced by the concatenation method may not be consistent [112]. The concatenation method treats every nucleotide of all

available genes equally and independently. This indicates that the genes having long sequence have more information about the species tree than the short genes. If the long genes happen to be incongruent with the true species tree, it may result in the wrong estimate of the species tree.

1.3.2 Conditional combination.

The conditional combination approach was proposed to overcome the restriction of the combination method that all genes must share the same histories. This approach performs a test to see if datasets have the same phylogenetic histories before combining them. Only the datasets with the same histories will be combined. Two currently used tests are the incongruence length difference test [29] and Templetons test [121].

1.3.3 Separation approach.

The separation approach estimates the gene trees separately for each genetic locus and uses the estimated gene trees to construct a species tree. This approach treats the gene tree as a character of the species tree and the gene trees, instead of DNA sequences, are thus used as the direct estimator of the species tree. Congruence among different gene trees provides strong information about the structure of the species tree. The consensus tree method is within the frame of the separation approach. The consensus tree method estimates the gene tree for each locus independently and assigns equal weight to each locus in order to combine the gene trees from different loci. However, all gene trees depend on the same species tree and are therefore not independent.

1. Gene tree parsimony. Gene tree parsimony is one of the most commonly used approaches to estimate the species tree. The idea of gene tree parsimony is quite straightforward. Given the gene trees, gene tree parsimony operates by finding the species tree or trees that minimize the number of hypothesized gene tree/species tree conflicts. Under the assumption that the observed gene tree/species tree conflicts are due to gene duplication, deep coalescence and horizontal transfer, the estimated species tree is the one that minimizes the weighted sum of these three events. The gene tree parsimony is implemented in the program GeneTree by Page [98]. It includes three steps.

   (a) The gene tree is inferred for each linkage partition.

   (b) Define the loss function. The loss function is the weighted sum of gene duplication/loss, deep coalescence and horizontal transfer. How to define the weights is still an open question.

   (c) The species tree is the one that minimizes the loss function.

   Weakness:

   (a) It does not take into account estimation errors of gene trees.

   (b) It may construct two optimal species trees.

2. Maximum likelihood method. Rannala and Yang (2003) [102] have derived the likelihood of the gene tree given the species tree using coalescent theory.

$$f(G|S) = \prod_{i=1}^{k}\{\prod_{j=n+1}^{m}[\frac{2}{\theta_i}\exp\{-\frac{j_i(j_i-1)}{\theta_i}t_{ij}\}] \times \exp\{-\frac{n_i(n_i-1)}{\theta_i}(\tau_i - \sum_{j=n+1}^{m} t_{ij})\}\}$$

If the gene trees are known, the species tree may be estimated by maximizing $f(G|S)$. Unfortunately, gene trees are generally unknown. The DNA sequences are instead the observed data. To estimate the species tree using DNA sequneces, we need one more step to derive the likelihood function $f(D|S)$.

$$f(D|S) = \int_G f(D|G)f(G|S)dG$$

$f(D|S)$ involves a formidable integration over gene trees. The tree search techniques used in the gene tree reconstruction method are not applicable here for estimating the species tree, because the function $f(D|S)$ requires the summation over all the possible gene trees. There is no technique currently available to estimate the species tree using maximum likelihood.

Alternatively, Bayesian techniques may be used to estimate the species tree. The Metropolis-Hastings sampler [84] [50] is often easier to carry out than maximizing $f(D|S)$. We will discuss the Bayesian approach in the next chapter.

CHAPTER 2

BAYESIAN HIERARCHICAL MODEL AND MARKOV CHAIN MONTE CARLO.

Traditional molecular-based phylogenetic analysis consists broadly of two steps: obtaining and aligning molecular sequences and inferring gene trees for those sequences. Under this paradigm, gene trees are generally considered to be synonymous with species trees, except when forces causing discordance between gene and species trees are obvious, such as horizontal gene transfer, deep coalescence, or gene duplication [79] [81]. Thus, in fact, phylogenetic analysis really consists of three elements: molecular sequences, gene trees and species trees. Identifying the relationships among these three elements and extracting useful information from each element are the key issues for constructing an appropriate model to explain the evolutionary history of a set of species.

One discussion in the literature revolves around which should be used as the direct estimator of the species tree, sequences or gene trees? Kluge and Wolf [68] [70] claim that natural data partitions do not exist and the species tree should be estimated using the whole sequence of the genome. They proposed a combined-data approach [68] [70] [95] in which the sequences from all available genes are concatenated into a single sequence, along with other phylogenetic characters such as morphology

or behavior. This method ignores the existence of the gene as the basic functional unit on the genome and treats nucleotides as the direct estimator of the species tree. This has drawn criticism [112] since it assumes that the longer the sequence the more precise the estimated species tree. It is now generally appreciated that gene trees in principle may not match the species tree irrespective of whether the gene has a long sequence or a short sequence. Indeed, recent work shows that under some combinations of branch lengths in the species tree, incongruent gene trees are more likely to occur than congruent gene trees [71] [17]. In other words, nucleotide or amino acid data are not consistent estimators of the species tree under some circumstances of speciation. Other approaches consider gene trees as the direct estimator of the species tree [98] [99]. This idea is based on Doyles [20] concept that nucleotides are characters of gene trees, whereas gene trees are characters of species trees [79]. This viewpoint suggests that using sequence data to infer species phylogeny requires two hierarchical levels of estimation: gene tree estimation and species tree estimation.

Methods for inferring gene trees from sequence data are numerous and have become extraordinarily sophisticated in recent years [21] [13]. However, methods for inferring species trees from gene trees are in their infancy, are not widely used and in general suffer from numerous statistical and methodological drawbacks. For example, the gene tree parsimony method [98] [110] and consensus tree methods [7] [15] [104] [57] both ignore the errors in gene tree estimation and generally assume that gene trees are estimated with perfect certainty. In this approach, maximum likelihood (ML) or maximum parsimony (MP) trees are built for each gene and used as the true gene trees to infer the species tree. Both methods then underestimate the variation in the procedure for inferring a species phylogeny. Moreover, some genes may be more

important than other genes in estimating species trees requiring a weighting that is difficult to incorporate into current methods.

Recent research has focused on the statistical model for species tree estimation. Here coalescent theory plays a central role in this model construction. Degnan and Salter [18] have derived the probability distribution for the topology of gene trees given the species tree. Slatkin and Pollack [111] specified a statistical model for the gene genealogies of two linked loci of three species. Coalescent theory has also been applied to forming the likelihood for genetic markers such as RFLPs, SNPs and AFLPs [93] [92].

The process for estimating gene trees and for estimating species trees should not be independent. Yet, when studying congruence among different genes in a phylogenetic study, gene trees are usually reconstructed for each gene independently. This assumption is questionable. Gene trees for different genes are dependent since they all depend on the species tree. For example, suppose we have data for nine genes from three species A, B and C. If the histories of the first 8 genes are (AB)C, it is clearly more likely for the last gene to have the (AB)C topology than any other topology. This is because the first 8 genes imply that the underlying species tree favors the generation of gene trees with the topology (AB)C. Consequently, it is more appropriate to assume only conditional independence of the gene trees given a common species tree. According to this assumption, the gene trees should then be estimated jointly across multiple loci.

These differences in the relevance of different gene trees to the estimation of the species tree suggests that a model for inferring the species tree using sequence data should have the following features:

1. It should simultaneously involve the distribution of sequences, gene trees and the species tree.

2. The underlying species tree should induce a marginal dependence in the gene trees which should then be inferred jointly across loci.

3. The model must take into account errors in the estimation of gene trees.

2.1 Bayesian hierarchical model.

In the equations that follow, we use the following abbreviations: D: Sequence data; G: a vector of gene trees; Λ: Parameters in the likelihood function except the gene tree vector G; S: Species trees; θ: Transformed effective population sizes.

The posterior probability of a species tree and θ is given by

$$f(S,\theta|D) = \frac{1}{f(D)} \int_{\Lambda} \int_{G} f(D|G,\Lambda) f(\Lambda) f(G|S,\theta) f(S) f(\theta) dG$$

.

Our Bayesian hierarchical structure consists of modeling the following components: $f(D|G,\Lambda)$, $f(\Lambda)$, $f(G|S,\theta)$, $f(S)$ and $f(\theta)$, each of which is explained below.

2.1.1 Likelihood $f(D|G,\Lambda)$.

Markovian models that assume independent sites dominate the likelihood based literature for both nucleotide and amino acid substitution [36]. It is worth mentioning that, while models for nucleotide and protein sequence data are the most common, our formulation allows for any type of underlying input data where $f(D|G,\Lambda)$ can be appropriately described. The quantity $f(D|G,\Lambda)$ will change according to the input

data and, for the same type of data, the most suitable model may be selected using a likelihood based model selection process [101] or information theory [85].

2.1.2 $f(\Lambda)$.

Λ includes the parameters in the substitution model and all other parameters in the likelihood function except the gene tree. Naturally, the prior on Λ will depend on the nature of the data at hand. For example, a variety of options for the prior of Λ are available in the Bayesian gene tree program, MrBayes [106].

2.1.3 $f(G|S, \theta)$.

The distribution of gene trees given species tree is derived from coalescent theory. Although the procedure can allow more general models, our initial implementation uses the coalescent theory in which random mating is assumed in each population. We also assume no gene flow after species divergences and no recombination within a locus but free recombination between loci.

The branch length in a species tree represents "time" (numbers of generations), whereas it is the expected number of mutations in a gene tree. To make the two parameters compatible, we transform to $\theta = 4N_e\mu$ where N_e is the effective population size and μ is the mutation rate measured as the expected number of nucleotide substitutions per site per generation.

The joint probability distribution of a gene tree topology and the $m-n$ coalescent times $t_{n+1}, \cdots, t_m$ for a single population reduced from m to n sampled individuals along a branch of length τ in a species tree was derived by Rannala and Yang (2003) to be:

$$exp-\frac{n(n-1)}{\theta}(\tau - \sum_{j=n+1}^{m} t_j) \prod_{j=n+1}^{m} (\frac{2}{\theta} exp(-\frac{j(j-1)}{\theta} t_j))$$

Thus $f(G|S,\theta)$ is the product of such probabilities across all the populations. For a vector of gene trees, G, that are independent given the species tree, we multiply these conditional likelihoods in turn to find $f(G|S,\theta)$. It should be noted that the species tree space is constrained because we assume that the gene split times of any two species predate their speciation time. The divergence times in the gene trees are always earlier than their counterparts in the species tree.

Note also that the θ may be different for different genes. For example, the mitochondrial and Y-chromosomal genes are uniparentally inherited and haploid. Thus in the data analyses in Chapter 4 below, we assume their effective population sizes are one-fourth that of autosomal markers.

2.1.4 $f(\theta)$.

We use independent gamma distributions as the priors of θ. The hyperparameters of the gamma distribution must also be chosen to be appropriate to the analysis.

2.1.5 $f(S)$.

We use a birth-and-death process [88] as the prior distribution of the species tree's topology and branch lengths. Given the speciation rate (s), extinction rate (e) and the number of species in study (n), the joint density of the topology (T) and branch lengths (τ) of a particular species tree is [128]:

$$f(T,\tau|n,\tau_1,s,e) = \frac{2^{n-1}}{n!(n-1)} \prod_{j=2}^{n-1} \frac{\lambda P_1(t_j)}{v_{t_1}}$$

where $v_{t_1} = 1 - \frac{1}{\rho}P(0,t_1)e^{(s-e)t_1}$, $P_1(t) = \frac{1}{\rho}P(0,t)^2 e^{(s-e)t_1}$.

and $P(0,t) = \frac{\rho(s-e)}{\rho s + (s(1-\rho)-e)e^{(s-e)t}}$.

Mutation rate variation among loci may influence the estimation of ancestral population sizes [129] [12]. If the ratios of rates between loci are known, we can incorporate them in the likelihood calculation [129]. We treat these relative mutation rates among loci as parameters in our model and assume that their prior follows the uniform (0,10) under the constraint that the average ratio is 1.

2.2 Properties of Likelihood function $f(G|S,\theta)$.

2.2.1 The likelihood function $f(G|S,\theta)$.

Let O_{ij} be the number of lineages of gene j leaving the population i, and let I_{ij} be the number of lineages of gene j entering the population i. Suppose there are N populations and M genes. The likelihood for the gene j and the population i is given by

$$[\prod_{k=O_{ij}+1}^{I_{ij}} \frac{2}{\theta_i}\exp(-\frac{k(k-1)}{\theta_i}t_{ij}^k)] \times \exp(-\frac{O_{ij}(O_{ij}-1)}{\theta_i}(\tau_i - \sum_{k=O_{ij}}^{I_{ij}} t_{ij}^k))$$

and $f(G|S,\theta)$ is the product of such likelihoods across all the populations and genes.

$$\prod_{j=1}^{M}\prod_{i=1}^{N}[\prod_{k=O_{ij}+1}^{I_{ij}} \frac{2}{\theta_i}\exp(-\frac{k(k-1)}{\theta_i}t_{ij}^k)] \times \exp(-\frac{O_{ij}(O_{ij}-1)}{\theta_i}(\tau_i - \sum_{k=O_{ij}}^{I_{ij}} t_{ij}^k))$$

The number of genes and populations are finite. It is valid to exchange the first two products of the likelihood function. Thus, $f(G|S,\theta)$ becomes

$$\prod_{i=1}^{N}\prod_{j=1}^{M}[\prod_{k=O_{ij}+1}^{I_{ij}} \frac{2}{\theta_i}\exp(-\frac{k(k-1)}{\theta_i}t_{ij}^k)] \times \exp(-\frac{O_{ij}(O_{ij}-1)}{\theta_i}(\tau_i - \sum_{k=O_{ij}}^{I_{ij}} t_{ij}^k))$$

The likelihood is the product of each population's likelihood.

$$f(G|S,\theta) = \prod_{i=1}^{N} L^i$$

.

Where L^i is the likelihood for population i.

2.2.2 Maximize $f(G|S,\theta)$.

Define $a_i(\tau_i) = \sum_{j=1}^{M} O_{ij}(O_{ij}-1)(\tau_i - \sum_{k=O_{ij}}^{I_{ij}} t_{ij}^k) + \sum_{k=O_{ij}+1}^{I_{ij}}(k(k-1)t_{ij}^k)$ and $b_i = \sum_{j=1}^{M}(I_{ij} - O_{ij})$. Then L^i can be rewritten in terms of $a_i(\tau_i)$ and b_i.

$$L^i = (\frac{2}{\theta_i})^{b_i} \exp(-\frac{a_i(\tau_i)}{\theta_i})$$

.

Note b_i is the total number of coalescences in the population i and $a_i(\tau_i)$ is the weighted sum of coalescence times in population i. It is obvious that $\sum_{i=1}^{N} b_i = TC$, where TC is the total number of coalescences across all the populations and $0 \leq b_i \leq TC$, $a_i(\tau_i) > 0$, $\theta_i > 4\mu$ because $\theta_i = 4N_e\mu$ and $N_e \geq 1$.

Lemma 2.2.1 *The likelihood function $f(G|S,\theta)$ is bounded.*

Proof It is clear that $f(G|S,\theta) > 0$.

$$\begin{aligned} L^i &= (\frac{2}{\theta_i})^{b_i} \exp(-\frac{a_i(\tau_i)}{\theta_i}) \\ &\leq (\frac{2}{\theta_i})^{b_i} \\ &\leq (\frac{2}{4\mu})^{b_i} \end{aligned}$$

If $2\mu > 1$, then $L^i < 1$, otherwise $L^i < (\frac{1}{2\mu})^{TC}$. We know that $f(G|S,\theta) = \prod_{i=1}^{N} L^i$ which is the product of bounded functions, indicating $f(G|S,\theta)$ itself is bounded. ∎

It is trivial to show that for any arbitrary b_i and $a_i(\tau_i)$, L^i is a concave function with respect to θ_i and $Max_{\theta_i}\{L^i\} = (\frac{2b_i}{a_i(\tau_i)})^{b_i} \exp(-2b_i)$ at $\theta_i = \frac{a_i(\tau_i)}{b_i}$. The values of b_i and $a(\tau_i)$ are fixed if and only if the topology and branch lengths of the species tree are fixed. If the species tree is given, the maximum likelihood estimator of θ_i is trivial. To maximize $f(G|S,\theta)$ with respect to b_i, $a_i(\tau_i)$ and θ_i simutaneously,

$$\begin{aligned}
\max_{b_i,a_i(\tau_i),\theta_i}\{f(G|S,\theta)\} &= max_{b_i,a_i(\tau_i),\theta_i}\{\prod_{i=1}^{N} L^i\} \\
&= max_{b_i,a_i(\tau_i),\theta}\{\prod_{i=1}^{N}(\frac{2}{\theta_i})^{b_i}\exp(-\frac{a_i(\tau_i)}{\theta_i})\} \\
&= max_{b_i,a_i(\tau_i)}\{\prod_{i=1}^{N}(\frac{2b_i}{a_i(\tau_i)})^{b_i}\exp(-b_i)\} \\
&= \exp(-TC)max_{b_i,a_i(\tau_i)}\{\prod_{i=1}^{N}(\frac{2b_i}{a_i(\tau_i)})^{b_i}\}
\end{aligned}$$

Therefore, we need to maximize $\prod_{i=1}^{N}(\frac{2b_i}{a_i(\tau_i)})^{b_i}$ with respect to b_i and $a_i(\tau_i)$ $\forall i = 1, \cdots N$. Note that $\sum_{i=1}^{M} b_i = TC$.

Let's consider a simple case when the θ_i's are equal. Then the likelihood function

$$\begin{aligned}
f(G|S,\theta) &= \prod_{i=1}^{N-1}(\frac{2}{\theta_i})^{b_i}\exp(-\frac{a_i(\tau_i)}{\theta_i}) \\
&= (\frac{2}{\theta})^{TC}\exp(-\frac{\sum a_i(\tau_i)}{\theta})
\end{aligned}$$

Now, $f(G|S,\theta)$ is maximized at $\theta = \frac{\sum a_i(\tau_i)}{TC}$ for which

$$f(G|S,\theta) = (\frac{2TC}{\sum a_i(\tau_i)})^{TC} \exp(-TC) \ (1).$$

Note (1) is a monotone decreasing function with respect to $\sum a_i(\tau_i)$. Thus, (1) is maximized at $\min\{\sum a_i(\tau_i)\}$.

Each node and its branch length τ in the species tree represents a population. We call this population the parent population and the two populations right below the parent population are the two daughter populations.

Lemma 2.2.2 *If the species tree topology is fixed and just node i moves up and down while the other nodes stay at the same place, the $a_i(\tau_i)$ of the parent population is monotone decreasing as node i goes up and increasing when the node i goes down. The $a_i(\tau_i)$s of the two daughter populations goes the opposite way.*

Proof The first coalescence time in the parent population decreases as the node i goes up, while the other coalescence times remain the same, leading to the increase of the $a_i(\tau_i)$ of the parent population. On the contrary, the gap between the speciation time and the last coalescence time in the two daughter populations increase as the node i goes up, resulting in the decrease of $a_i(\tau_i)$ of the two daughter population. ∎

Lemma 2.2.3 *If node i moves up and down while the other nodes are fixed and θ_i's are equal, then the $\sum_{i=1}^{N} a_i(\tau_i)$ is continuous and monotone decreasing with respect to the branch length τ_i of node i. If θ_i's are not equal, $\sum_{i=1}^{N} a_i(\tau_i)$ is not continuous and has jumps at each coalescence time.*

Proof To prove the first conclusion, consider a species tree with fixed topology in Figure 2.1. The node i moves among the coalescence times T_1, T_2 and T_3 while keeping the other nodes fixed. There are two possible ways to move the node i.

(a)The node i may move between two adjacent coalescent times.

(b)The node i may move across a coalescent time.

For the case (a), it is trivial to show that $\sum_{i=1}^{N} a_i(\tau_i)$ is continuous because it is a weighted sum of the continuous functions with fixed weights. For the case (b), it suffices to show that $\lim_{node\uparrow T_1} \sum_{i=1}^{N} a_i(\tau_i) = lim_{node\downarrow T_1} \sum_{i=1}^{N} a_i(\tau_i)$.

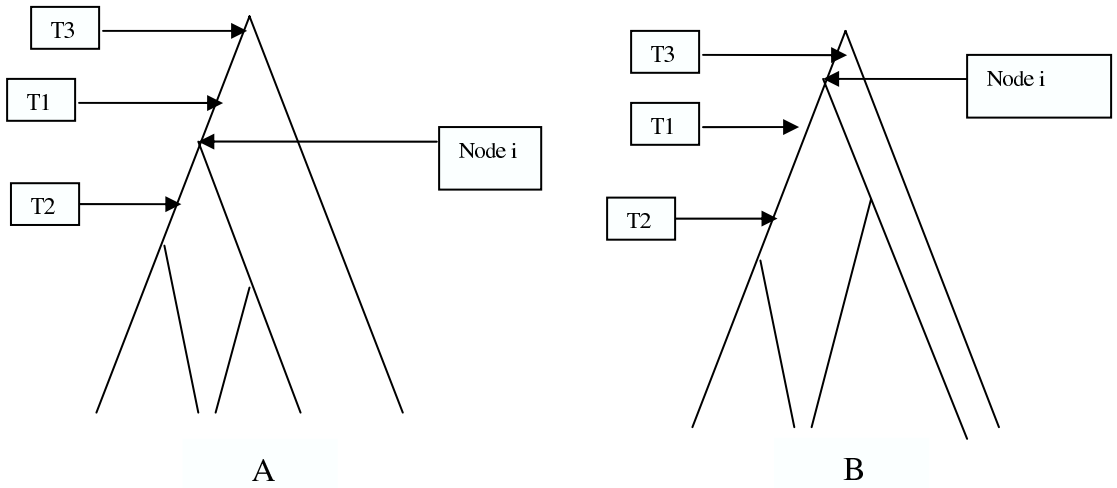

Figure 2.1: (A) T_1, T_3 are the first two coalescent times in the population i, while T_2 is the last coalescence time in the left daughter population. The node i moves up towards to T_1.(B) T_3 is the first coalescent time in the population i, while T_1 and T_2 are the last two coalescence times in the left daughter population. The node i moves up towards to T_1.

Let's consider the left side of the equation first (Figure 2.1 A). Let a_1 be the $a_i(\tau_i)$ of the population i, a_2 be the $a_i(\tau_i)$ of the left daughter population and a_3 be the $a_i(\tau_i)$ of the right daughter population. $\sum_{i=1}^{N} a_i(\tau_i) = a_1 + a_2 + a_3 + C$. C is a constant since the other nodes in the species tree are fixed. Moreover, the coalescence time T_1 belongs to either the left daughter population or the right daughter population

when node i crosses it. We assume it belongs to the left daughter population. The proof is the same if it belongs to the right daughter. We need to show that $a_1 + a_2$ is continuous with respect τ_i, ie., $\lim_{node\uparrow T_1}(a_1 + a_2) = lim_{node\downarrow T_1}(a_1 + a_2)$.

$$
\begin{aligned}
a_1 &= \sum_{j=1}^{N} I_{1j}(I_{1j}-1)(T_1-\tau) + (I_{1j}-1)(I_{1j}-2)(T_3-T_1) + C_1 \\
a_2 &= \sum_{j=1}^{N} O_{2j}(O_{2j}-1)(\tau - T2) + C_2 \\
\lim_{node\uparrow T_1} a_1 &= \sum_{j=1}^{N} (I_{1j}-1)(I_{1j}-2)(T_3-T_1) + C_1 \\
\lim_{node\uparrow T_1} a_2 &= \sum_{j=1}^{N} O_{2j}(O_{2j}-1)(T_1-T_2) + C_2
\end{aligned}
$$

$$
\Rightarrow \lim_{node\uparrow T1} a_1 + a_2 = \sum_{genetrees} O_{2j}(O_{2j}-1)(T1-T2) + (I_{1j}-1)(I_{1j}-2)(T3-T1) + C_1 + C_2
$$

If node i moves downwards to T_1, the number of coalescence times in the population i will decrease by 1, whereas the number of coalescence times of the left daughter population will increase by 1.

$$
\begin{aligned}
a_1 &= \sum_{j=1}^{N} (I_{1j}-1)(I_{1j}-2)(T3-\tau) + C_1 \\
a_2 &= \sum_{j=1}^{N} O_{2j}(O_{2j}-1)(T1-T2) + (O_{2j}-1)(O_{2j}-2)(\tau - T1) + C_2 \\
\lim_{node\downarrow T_1} a_1 &= \sum_{j=1}^{N} (I_{1j}-1)(I_{1j}-2)(T3-T1) + C_1 \\
\lim_{node\downarrow T_1} a_2 &= \sum_{j=1}^{N} O_{2j}(O_{2j}+1)(T1-T2) + C_2
\end{aligned}
$$

$$\Rightarrow \lim_{node\downarrow T_1} a_1 + a_2 = \sum_{j=1}^{N} O_{2j}(O_{2j}-1)(T1-T2) + (I_{1j}-1)(I_{1j}-2)(T3-T1) + C_1 + C_2$$

Therefore, $\lim_{node\uparrow T_1} a_1 + a_2 = lim_{node\downarrow T_1} a_1 + a_2$, which also implies that

$$\lim_{node\uparrow T_1} \sum_{i=1}^{N} a_i(\tau_i) = lim_{node\downarrow T_1} \sum_{i=1}^{N} a_i(\tau_i)$$

We have shown that $\sum_{i=1}^{N-1} a_i(\tau_i)$ is continuous at each coalescent time for a fixed topology in which node i will never go beyond its father node or its two daughter nodes. What if node i goes beyond its father node j? It turns out that $\sum_{i=1}^{N} a_i(\tau_i)$ is still continuous. Denote the divergence time of node j by τ_f. We need to show

$$\lim_{node\uparrow \tau_f} \sum_{i=1}^{N} a_i(\tau_i) = lim_{node\downarrow T_1} \sum_{i=1}^{N} a_i(\tau_i)$$

.

The $\sum_{i=1}^{N} a_i(\tau_i)$ does not depend on the order of the populations. So $lim_{node\rightarrow T_f} \sum_{i=1}^{N-1} a_i(\tau_i)$ remains the same no matter node i goes down to T_f or goes up to T_f. So it is true that

$$\lim_{node\uparrow \tau_f} \sum_{i=1}^{N} a_i(\tau_i) = lim_{node\downarrow T_f} \sum_{i=1}^{N} a_i(\tau_i)$$

.

The proof of continuity of $\sum_{i=1}^{N} a_i(\tau_i)$ is complete.

Let's prove that $\sum_{i=1}^{N} a_i(\tau_i)$ is monotone decreasing with respect to τ_i. Let a_{new} be the $\sum_{i=1}^{N} a_i(\tau_i)$ after node i moves up and a_{old} be the $\sum_{i=1}^{N} a_i(\tau_i)$ before node i moves up. Suppose τ_i increases by δt. Let I_1 be the number of lineages entering into the parent population and t_1 be the difference between the first coalescence time and the

divergence time τ_i of the parent population. Let O_2 be the number of genes leaving the left daughter population and t_2 be the difference between the divergence time τ_i and the last coalescence time of the left daughter population. Let O_3 be the number of genes leaving the right daughter population and t_3 be the difference between the divergence time τ_i and the last coalescence time of the right daughter population. Note $O_2 > 0$ and $O_3 > 0$. Also, note that $I_1 = O_2 + O_3$. The difference between the a_{new} and a_{old} is

$$I_1(I_1-1)(t_1-\delta t)+O_2(O_2-1)(t_2+\delta t)+O_3(O_3-1)(t_3+\delta t)-[I_1(I_1-1)t_1+O_2(O_2-1)t_2+O_3(O_3-1)t_3]$$

The other terms are canceled. This quantity can be simplified as

$$\begin{aligned} a_{new} - a_{old} &= \delta t \times [O_2(O_2-1) + O_3(O_3-1) - I_1(I_1-1)] \\ &= -2O_2O_3\delta t \\ &< 0 \end{aligned}$$

We have proved that $\sum_{i=1}^{N} a_i(\tau_i)$ is monotone decreasing with respect to τ_i.

If θ_i's are not equal, $\sum_{i=1}^{N} a_i(\tau_i)$ is still continuous between two adjancent coalescence times, because $\sum_{i=1}^{N} a_i(\tau_i)$ is a weighted sum of continuous functions with fixed weights, but it has jumps at each coalescent time, because

$$\lim_{node\uparrow t} Max_\theta\{L(p)\} \neq lim_{node\downarrow t} Max_\theta\{L(p)\}.$$

The second statement is proved. ∎

The space of the species tree is restricted if the gene trees are given, because it is assumed that the gene splits predate the speciation times. Let Ω be the restricted

pair of species	gene split time for gene1	gene split time for gene2	minimum
A,B	0.1	0.12	0.1
A,C	0.1	0.07	0.07
A,D	0.1	0.09	0.09
B,C	0.05	0.12	0.05
B,D	0.06	0.12	0.06
C,D	0.06	0.09	0.06

Table 2.1: The gene split times for each pair of species.

species tree space and G be a vector of gene trees. The assumption says that given G, $S < G \; \forall \; S \in \Omega$. Let's use an example to illustrate how to find Ω. Let gt(A,B) be the gene split time for the species A and B. Let st(A,B) be the speciation time of A and B.

Suppose there are four species, A, B, C, and D. and two genes (a very simple case), gene1 and gene2. The gene trees for the two genes are (A:0.1, (D:0.06, (B:0.05 ,C:0.05))), (B:0.12, (D:0.09, (A:0.07,C:0.07))). Note that both species trees and gene trees are clock trees. The gene split times for the two genes are listed in Table 2.1

The gene split time of species A and B is 0.1 for gene1 and 0.12 for gene2. Their speciation time should be less than any of these two gene split times. Consequently, their speciation time is less than the minimum of the two gene split times which is 0.1. Similarly, the speciation time of A and C must be less than 0.07 and the speciation time of B and C must be less than 0.05. The space of species tree Ω must satisfy the following constraints, $st(A,B) < 0.1,\ st(A,C) < 0.07,\ st(A,D) < 0.09,\ st(C,B) < 0.05,\ st(B,D) < 0.06,\ st(C,D) < 0.06$. Generally, there are $\begin{pmatrix} n \\ 2 \end{pmatrix}$ constraints for n species. However, some constraints are redundant. For example,

if $st(C, B) < 0.05$ *and* $st(A, C) < 0.07$, then st(A,B) is less than 0.07. So the first constraint $st(A, B) < 0.1$ is not necessary.

The following lemma shows that We need only $(n-1)$ instead of $\binom{n}{2}$ constraints for n species.

Lemma 2.2.4 *If there are n species, then given the gene tree vector G, the minimum number of constraints on the speciation times required to be equivalent to the* $\binom{n}{2}$ *constraints simultaneously is* $n - 1$.

Proof To prove the lemma, we will construct $(n - 1)$ constraints (we call them the minimum constraints) and show these constraints are sufficient and necessary. The $(n - 1)$ minimum constraints are constructed as follows.

step1: Choose the minimum of the $\binom{n}{2}$ constraints as the first constraint. We call it "constraint1" and the two species involved in the constraint1 are S1 and S2.

step2: Choose the minimum of the $\binom{n}{2}$ constraints in which one species is either S1 or S2 and the other one is a new species S3. This is the constraint2.

step3: Choose the minimum of the $\binom{n}{2}$ constraints in which one species is S1, S2 or S3 and the other one is a new species S4. This is the constraint3.

step4: Repeat until all the species are in the constraints.

Let's use an example to illustrate how to construct the $(n-1)$ minimum constraints. For the previous example of four species, there are six constraints: $st(A, B) < 0.1$, $st(A, C) < 0.07$, $st(A, D) < 0.09$, $st(C, B) < 0.05$, $st(B, D) < 0.06$, $st(C, D) < 0.06$.

step1: The minimum of the six constraints is $st(C, B) < 0.05$ and the two species in the constraint are C and B.

step2: There are three constraints in which one species is either C or B and the other is a new species. The constraints are $st(A, B) < 0.1$, $st(A, C) < 0.07$,

$st(B, D) < 0.06$, $st(C, D) < 0.06$. The minimum is $st(B, D) < 0.06$ or $st(C, D) < 0.06$. The new species in both constraints are the same, species D. In this situation, the two constraints are equivalent. We can choose either of them. Let's say we choose $st(B, D) < 0.06$ as constraint 2.

step3: There are three constraints in which one species is C, B or D and the other is a new species. The constraints are $st(A, B) < 0.1$, $st(A, C) < 0.07$, $st(A, D) < 0.09$. The minimum is $st(A, C) < 0.07$.

step4: We already have all the species in the constraints. So we stop.

The $(n-1)$ constraints we constructed for the previous example are $st(C, B) < 0.05$, $st(B, D) < 0.06$, $st(A, C) < 0.07$.

The necessary part of the lemma is trivial since the $(n-1)$ minimum constraints are within the $\begin{pmatrix} n \\ 2 \end{pmatrix}$ constraints.

To prove the sufficiency, we want to show the $(n-1)$ minimum constraints constructed in step1 - 4 are sufficient. In each step, a new species is introduced into the minimum constraints. The constraints other than the minimum constraints relating to the species already within the minimum constraints are automatically satisfied since we always choose the minimum constraint for the new species. After step4 when all the species have been within the minimum constraints, the $\begin{pmatrix} n \\ 2 \end{pmatrix}$ constraints will all be automatically satisfied. ∎

The $(n-1)$ minimum constraints can be represented by a tree—the Maximum Tree since it represents the maximum species tree in the restricted space. For example, the minimum constraints of the previous example constructs the following Maximum Tree (((B:0.05,C:0.05),D:0.06),A:0.07).

Theorem 2.2.5 *The global MLE of species tree exists and it is the Maximum Tree (MT) if the θ_i's are equal.*

Proof By Lemma 2.2.1, the likelihood function $f(G|S,\theta)$ is bounded, indicating that the maximum of $f(G|S,\theta)$ with respect to the species tree S exists.

We have proved in the Lemma 2.2.1 that if θ_i's are equal, the $f(G|S,\theta)$ is maximized at $\min\{\sum a_i(\tau_i)\}$. By Lemma 2.2.3, $\sum a_i(\tau_i)$ is continuously increasing as any arbitrary node i moves up. Furthermore, the Maximum Tree is the largest tree and $\min\{\sum a_i(\tau_i)\}$ is achieved when the species tree reaches the Maximum Tree. This completes the proof. ∎

The special case of Theorem 2.2.5 is that there is only one gene tree. Then the Maximum Tree is identical to the gene tree. In this case, the MLE of the species tree is the gene tree.

Theorem 2.2.6 *For any fixed set of species, the Maximum Tree is a consistent estimator of the species tree.*

Proof Let S be the true species tree and n be the number of genes. Suppose there are N populations in the species tree. By definition, MT is consistent if MT converges to S in probability. It suffices to show that all the divergence times in MT converge to their counterparts in S. Let MT_i be the divergence time of an arbitrary population i in MT and S_i be its counterpart in S. We want to show that as $n \to \infty$,

$$Prob[|MT_i - S_i| > \epsilon] \to 0 \ \forall \ \epsilon > 0.$$

By defination, MT_i is the minimum of the divergence times of population i across all gene trees. Let t_{ij} be the divergence time of population i for the gene j. Then

$MT_i = \min\{t_i, i = 1 \cdots n\}$. t_i is a simple random sample from a distribution F derived from coalescent theory. Note that $MT_i \geq S_i$.

$$Prob(|MT_i - S_i| > \epsilon) = Prob(MT_i - S_i > \epsilon) = (1 - Prob(t_i - S_i > \epsilon))^n \ (*)$$

For any $\epsilon > 0$, $(*) \to 0$ as $n \to \infty$. The proof applies to any population in MT. Therefore, we have shown that

$$Prob[|MT_i - S_i| > \epsilon] \to 0 \ as \ n \to \infty \ \forall \ \epsilon > 0 \ \forall \ i = 1 \cdots N.$$

By definition, MT is consistent if MT converges to S in probability. We need to show that as $n \to \infty$

$$Prob[|MT - S| > \epsilon] \to 0 \ \forall \ \epsilon > 0.$$

Define $|MT - S|$ as $\max\{|MT_i - S_i|, i = 1, \cdots, Ni\}$, Then

$$\begin{aligned} Prob[|MT - S| > \epsilon] &= Prob[\max\{|MT_i - S_i|\} > \epsilon] \\ &= \max\{(1 - Prob(t_i - S_i > \epsilon))^n\} \\ &\to 0 \end{aligned}$$

The maximum can be taken out of the Prob[.] because the number of populations is finite. The proof is complete. ∎

2.2.3 Maximum likelihood estimate (MLE).

If the θ's are not equal, it is difficult to derive the theoretical solution for the MLE. If the gene trees are known, any tree search algorithm for estimating the gene

phylogeny can be adopted here to search for the MLE of the species tree. If the gene trees are unknown, the likelihood function becomes

$$f(D|S,\theta) = \int_G \int_\Lambda f(D|G,\Lambda) f(\Lambda) f(G|S,\theta) dG d\Lambda$$

Apparently, the theoretical solution of the maximum is intractable. Alternatively, a numerical method like MCMC may be implemented to maximize $f(D|S,\theta)$ with respect to S and θ. We have shown that $\hat{\theta}$ can be easily derived if S is given.

$$\begin{aligned}
f(D|S,\hat{\theta}) &= \int_G \int_\Lambda f(D|G,\Lambda) f(\Lambda) f(G|S,\hat{\theta}) dG d\Lambda \\
&= \int_G \int_\Lambda f(D|G,\Lambda) f(\Lambda) \frac{f(G)}{f(G)} f(G|S,\hat{\theta}) dG d\Lambda \\
&= f(D) \int_G \int_\Lambda f(G,\Lambda|D) \frac{f(G|S,\hat{\theta})}{f(G)} dG d\Lambda \\
&= f(D) \int_G f(G|D) \frac{f(G|S,\hat{\theta})}{f(G)} dG \\
&= \frac{1}{N_1} \sum_{i=1}^{N_1} \frac{f(G_i|S,\hat{\theta})}{f(G_i)}
\end{aligned}$$

Thus, for each species tree, $f(D|S,\hat{\theta})$ can be approximated by the average of $\frac{f(G_i|S,\hat{\theta})}{f(G_i)}$ where G_i is the sample from the posterior distribution $f(G|D)$. The algorithm is

1. Generate a sample from the posterior distribution of gene trees using MrBayes. Calculate and save the prior of gene trees $f(G_i)$ in order to calculate the ratio $\frac{f(G_i|S,\hat{\theta})}{f(G_i)}$.

2. Add the new likelihood function $f(D|S,\hat{\theta})$ into PAUP* or PHYLIP and calculate $f(G_i|S,\hat{\theta})$ for each G_i from the sample of gene trees. $f(D|S,\hat{\theta})$ is approximated by $\frac{1}{N_1}\sum_{i=1}^{N_1}\frac{f(G_i|S,\hat{\theta})}{f(G_i)}$. Use the tree search algorithm in PAUP* or PHYLIP to find the optimal species trees.

$f(G)$ is crucial to the efficiency of the algorithm. It must be close to $f(G|\hat{S},\hat{\theta})$ in which $\hat{S}$ is the MLE. I suggest using $f(G|MT,\hat{\theta})$ as the prior to generate gene trees.

2.3 Markov Chain Monte Carlo (MCMC).

The entire species tree estimation procedure consists of three steps.

Step1 (within MrBayes): Generate vectors of gene trees from MrBayes using the approximate prior, K(G), based on a Maximum species tree estimate in the Hastings ratio to decide on acceptance of each vector into the Markov chain.

Step2: Using a second MCMC algorithm, generate species trees from the distribution compatible with the gene trees given by the approximate posterior distribution $K(G|D)$ from step 1.

Step3: Use importance sampling to align the results with what would have occurred if the initial sample had been from the true prior, f(G).

Markov Chain Monte Carlo (MCMC) is implemented to evaluate the posterior distribution of the species tree since f(D) involves an intractable integral. The posterior distribution of the species tree can be formulated as follows,

$$
\begin{aligned}
f(S,\theta|D) &= \int_\Lambda \int_G f(S,\theta,G,\Lambda|D) dG d\Lambda \\
&= \frac{1}{f(D)} \int_\Lambda \int_G f(D|G,\Lambda) f(\Lambda) f(G|S,\theta) f(\theta) f(S) dG d\Lambda \\
&= \frac{1}{f(D)} \int_\Lambda \int_G f(D|G,\Lambda) f(\Lambda) \frac{f(G)}{f(G)} f(G|S,theta) f(\theta) f(S) dG d\Lambda \\
&= \frac{1}{f(D)} \int_\Lambda \int_G f(D|G,\Lambda) f(\Lambda) f(G) \frac{f(G|S,\theta) f(\theta) f(S)}{f(G)} dG d\Lambda \\
&= \int_G f(G|D) f(S,\theta|G) dG \quad (1)
\end{aligned}
$$

The posterior of species tree and population sizes given data, $f(S,\theta|D)$, is the posterior of species tree and θ given gene trees $f(S,\theta|G)$ weighted by $f(G|D)$. This motivates our algorithm to generate the posterior distribution of gene trees first and then use these gene trees to generate the posterior for the species tree.

However, in the first stage of using DNA sequences to estimate the posterior of gene trees, the prior of gene trees, f(G), is unknown. Theoretically, f(G) is equal to the integral of $f(G|S)$ with respect to the species tree (topology and branch lengths) and population size θ, namely,

$$
f(G) = \int_\theta \int_S f(G|S,\theta) f(S) f(\theta) dS d\theta.
$$

It is by no means trivial to calculate f(G). Instead we use an approximation to this prior under the assumption that the gene splitting time is earlier than the speciation time, within MrBayes to define the Markov chain. In particular, for a given vector of gene trees we form the "maximum tree" defined earlier as the ultrametric tree that has the maximum divergence times for a species tree that is compatible with all the gene trees in the vector. We then apply Rannala and Yang's formula

(page 29) for the distribution of gene tree topologies consistent with this maximum tree to find an approximate prior K(G). Here, the integral with respect to the population sizes, θ, is approximated using the Monte Carlo method. This prior is used to define the Hastings ratios in MrBayes that decide whether a vector of gene trees is accepted into the Markov chain. The chain is then run to convergence generating a sample from the approximate posterior distribution $K(G|D)$. We save a subsample from this chain $G_1, G_2, \cdots, G_N$ along with the associated approximate priors $K(G_1), K(G_2), \cdots, K(G_N)$ to be used in steps 2 and 3.

In step 2 we find the posterior distribution of the species tree given the gene tree vectors generated in step 1. Here a second MCMC algorithm is applied. For this algorithm, the birth-and-death process is used to define the prior distribution of species trees (Nee, May, and Harvey, 1994) [88] and the likelihood is again defined by coalescent theory via Rannala and Yang's formula. The movement strategy employs a random selection of nodes and replacement uniformly within a random band that maintained the constraints while adjusting the topology where needed.

Step 2 provides k samples from each of the gene tree vectors $G_1, G_2, \cdots, G_N$ arising from the samples in step 1. Finally, importance sampling is applied to find the posterior distribution of species trees given the data. Note that:

$$\begin{aligned} f(S, \theta|D) &= \int_G f(G|D) f(S, \theta|G) dG \\ &= \int_G K(G|D) \frac{f(G)}{K(G)} f(S, \theta|G) dG \end{aligned}$$

The i^{th} sample from step 1 gave the value for $K(G_i|D)$. We need to multiply this by $\frac{f(G_i)}{K(G_i)}$ in order to align it with $f(G_i|D)$ and produce the samples from the

true posterior. $f(G_i)$ is not known but we can apply the harmonic mean technique (Newton and Raftery, 1994) [91] to estimate it. In particular, we have

$$\frac{1}{f(G_i)} \propto \int_S \frac{1}{f(G_i)} f(S)dS \;\; = \int_S \frac{1}{f(G_i|S)} \frac{f(G_i|S)}{f(G_i)} f(S)dS = \int_S \frac{1}{f(G_i|S)} f(S|G_i)dS$$

where the constant of proportionality, α, is the probability that a random species tree chosen from the birth-and-death model satisfies the constraints associated with G_i. Thus, a consistent estimate of $f(G_i)$ is given by

$$\hat{f}(G_i) = \hat{\alpha_i}(\sum_{j=1}^{k} \frac{1}{f(G_i|S_j)})^{-1}$$

using the samples $S_1, \cdots, S_k$ from $f(S|G_i)$ found in step 2. The value of $\hat{\alpha_i}$ is found by averaging $f(S)$ over randomly sampled trees from the constraint space induced by G_i. The final sample from the joint posterior distribution of S and G given D are then the pairs $\{(S_1, G_1), (S_2, G_1), \cdots, (S_k, G_1)\}, \cdots, \{(S_1, G_N), (S_2, G_N), \cdots, (S_k, G_N)\}$ where the block of pairs $\{(S_1, G_i), (S_2, G_i), \cdots, (S_k, G_i)\}$ is given total weight

$$\frac{\hat{f}G_i}{K(G_i)}(\sum_{i=1}^{N} \frac{\hat{f}G_i}{K(G_i)})^{-1}.$$

The source codes (written in the C language) of the revised MrBayes and BEST can be downloaded at www.stat.ohio-state.edu/~ dkp/BEST. You can also find the description and sample control files for the two programs.

2.4 Comparison with the Bayesian concatenation method and Bayesian consensus tree method.

In this section, we will compare our method with the concatenation method, and with the consensus tree method using Bayesian techniques. The Bayesian concatenation method (BCM) refers to the concatenation method using Bayesian approaches to infer gene trees [97]. Similarly, the Bayesian consensus tree method (BCT) estimates the posterior distribution of trees separately for each gene and the resulting gene trees for each gene are then pooled together as the posterior distribution of species trees [4]. The consensus tree of the posterior is then used as the point estimate summary of this species tree distribution.

Let G_i be the gene trees for gene i, D_i be the DNA sequences for gene i and take the number of genes to be k. The model of the Bayesian consensus tree method is straightforward because it assumes independent loci. The likelihood of DNA sequences given gene trees is just the product of the likelihood for each gene.

$$L^{BCT} = f(D|G) = f(D_1 \cdots D_k|G_1 \cdots G_k) = \prod_{i=1}^{k} f(D_i|G_i)$$

The prior of the gene trees for different genes is the product of the prior for each gene.

$$Prior^{BCT} = f(G) = f(G_1 \cdots G_k) = \prod_{i=1}^{k} f(G_i)$$

For the Bayesian concatenation method, the likelihood is a little different since all the genes should have the same tree G^*.

$$L^{BCM} = f(D_1 \cdots D_k | G^*) = \prod_{i=1}^{k} f(D_i | G^*)$$

The prior of gene trees here assumes that the gene trees from k genes are all the same.

$$Prior^{BCM} = f(G^*)$$

Comparing the likelihoods of the two methods, it is clear that L^{BCT} has more parameters than L^{BCM} because genes can take different trees in the Bayesian consensus tree method, whereas genes are typically assumed to follow the same tree in the Bayesian concatenation method. The parameter space is constrained for the Bayesian concatenation method. Consequently, the Bayesian consensus tree method will always provide a better fit of model to data but possibly at the expense of introducing extra variability.

The priors of the two methods are also different in another respect. The Bayesian consensus tree method uses independent gene tree priors, whereas the Bayesian concatenation method uses a joint prior in which the gene trees across k genes are correlated with correlation = 1 (because it assumes that all the gene trees are the same tree). The independent prior implies not only that the gene trees are themselves independent but also that the gene trees and species trees are independent. (It is possible that there is a single gene tree that depends on the species tree and that the others are independent of the species tree. But this does not change the following arguments). If gene trees and species trees are independent, then the gene trees would provide no value for inferring species trees. The independent prior is then valid only if we assume that gene trees and species trees are identical, which is not always true.

Thus, the joint prior appears to be more appropriate than the independent prior if we assume the species tree exists and is distinct from gene trees.

Although the trees estimated by the concatenation method and the consensus tree method are treated as species trees, they are actually gene trees. Theory does not guarantee that such estimated gene trees will be close to the species tree and can thereby be used as the estimate of the species tree [71] [17]. But it is clear that neither method facilitates estimation of important parameters in the evolutionary history of species such as population sizes or speciation times. Of course, speciation times here are distinct from gene divergence times due to the coalescent process [23].

In the continuum from concatenation to consensus methods, the technique proposed here is an intermediate approach that takes advantage of both methods. Firstly, the likelihood portion of our method is much like the one in the consensus tree method, because we allow the genes to have different trees. Secondly we use the joint prior instead of the independent prior for gene trees. But we do not assume the correlation=1. Instead, we use coalescent theory to specify the correlation structure among gene trees. After having generated samples from the posterior of gene trees for each gene, coalescent theory is used to combine those gene trees to infer the species tree. By choosing a particular prior of the species tree and the distribution of gene trees given the species tree, the Bayesian hierarchical model can be reduced to the Bayesian concatenation method or consensus tree method as special cases. For example, let the distribution of gene trees given species trees $f(G|S)$ be a degenerate distribution with all gene trees and the species tree always equal. In this case, estimating the species trees S is equivalent to estimating gene trees G. The posterior $f(S|D)$ in the

Bayesian hierarchical model would then be equal to the posterior $f(G^*|D)$ in the Bayesian concatenation method.

CHAPTER 3

SIMULATION STUDY

The statistical properties of our Bayesian hierarchical model are explored through simulation in this chapter. Reasonable assumptions are necessary to simplify a complicated reality that consists of not only the truth but also the random noise. However, it may cause serious problems if the model is over-simplified and the assumptions in the model are not even close to the reality. Biological justification is one of the most important aspects for any statistical models used to explain biological process. The justification of assumptions should then be a collaborative work of biologists and statisticians. We have already discussed some assumptions of our model such as the use of the joint distribution of gene trees. In this chapter, we concentrate on the statistical properties of the new estimation technique by assuming the model is true.

3.1 Goal of the simulation study.

The simulation procedure has two consecutive steps. Gene trees are first generated from a pre-specified species tree using coalescent theory. A substitution model is then assigned to each gene tree. The DNA sequences are simulated from the gene trees with the corresponding substitution model.

The data are analyzed using our method to examine if the method can deliver reasonable estimates of parameters in the model. The primary goal of this simulation study is to understand

1. The effect of the number of genes on the posterior probability of the true species tree.

2. The effect of the proportion of gene trees matching the true species tree on the posterior probability of the true species tree.

3. The comparison of our method with the concatenation method and the consensus tree method.

3.2 Methodology.

A tree with speciation times and population sizes is specified as the true species tree. We used MCMCcoal [130] to generate genealogical trees from the pre-specified species tree. We did two analyses. In the first analysis, the proportion of gene trees matching the true species tree was controlled to be a small number, while in the second analysis the proportion was relatively high. For the first analysis, the data set of 4 and 8 species were generated using the following true species phylogeny.

True species phylogeny for 4 species (species sa, sb, sc, and sd):

(((sa , sb) : 0.0057 #.008, sc) : 0.0062 #.005, sd) :.014 #.006;

True species phylogeny for 8 species (species sa, sb, sc, sd, se, sf, sg, and sh):

((((sa, sb) : 0.005 #.008, sc) : 0.0076 #.009, sd) :.008 #.005,

((se,sf):.003 #.001,(sg,sh):0.0068#0.014):0.007#0.001):0.018#0.011;

For the second analysis, the true species trees were:

4 species: (((H , C) : 0.0057 #.005, G) : 0.0102 #.005, O) :.024 #.006;

8 species: ((((sa, sb) : 0.003 #.005, sc) : 0.0086 #.004, sd) :.014 #.002,

((se,sf):.003 #.001,(sg,sh):0.0048#0.004):0.012#0.001):0.018#0.021;

Following Yang [130], the numbers after the colons are the speciation times in units of substitutions per site. The numbers after the pound signs (#) are the ancestral effective population sizes in units of substitutions per site.

We simulated 20, 40, 80 and 120 gene trees for each true species tree. These gene trees were then used to estimate the species tree using two exponential distributions with mean of 0.005 and 0.001 as the prior of θ which was the transformed effective population sizes.

3.3 Result and Discussion.

3.3.1 The number of genes vs. the posterior probability of the true species tree.

Under the scenarios in which the proportion of gene trees matching the species tree was high, we found that the correct species tree might be recovered with high probability with fewer than 3 genes (Figure 3.1A). However, if the proportion of gene trees matching the species tree was low, we found that at least 120 genes were required to accurately estimate the species tree in the case of 8 species (Figure 3.1B and C). Remarkably, the method was able to correctly reconstruct the species tree with high probability even when the proportion of gene trees matching the species tree was less than 10% (Figure 3.1C).

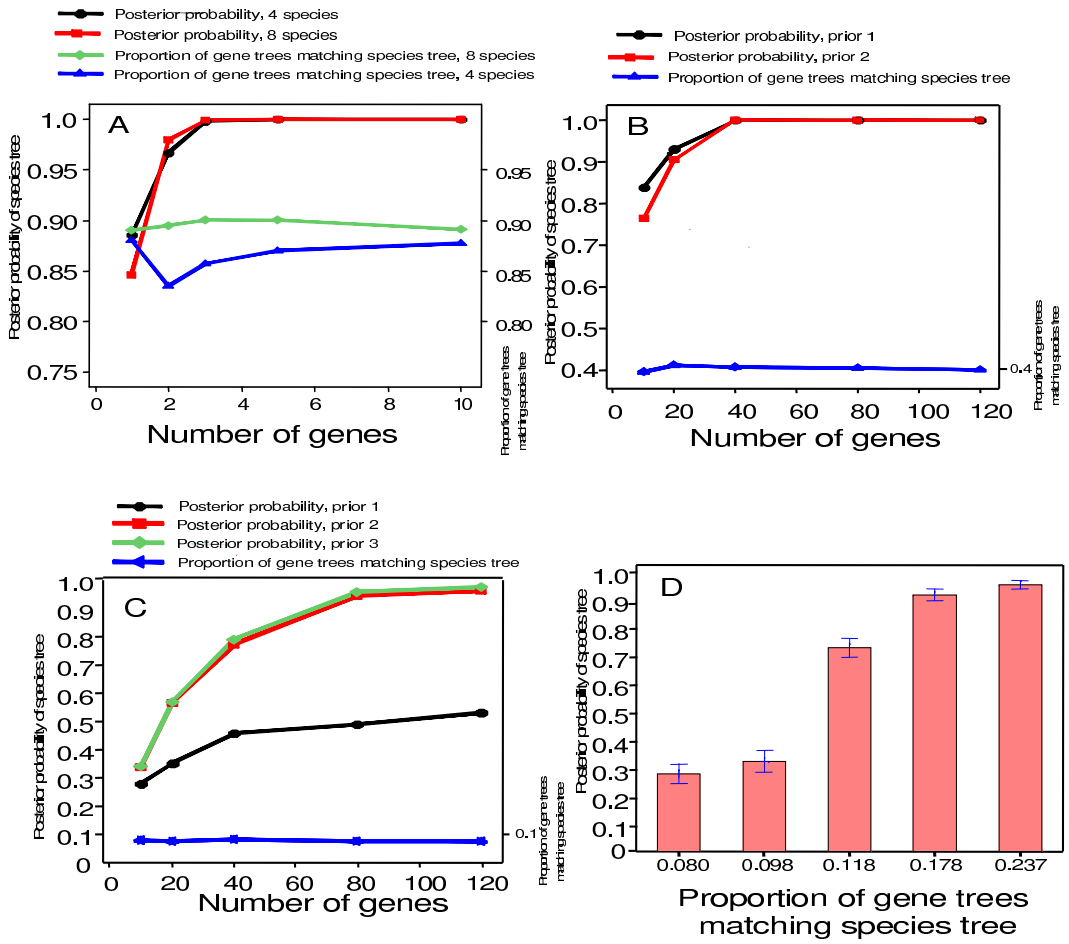

Figure 3.1: Robustness and efficiency of the joint prior method for estimating species trees. A) The number of gene required to resolve the correct species tree with 4 and 8 species when the proportion of gene trees matching the species tree (P) is high. Here P varies between 83% and 90% (in blue and green, 100 gene trees per simulation) because the critical internodes in the species tree are relatively long on the scale of the effective population size (θ). The gamma-distributed prior on ancestral population size for each node was (1,200). B. The number of gene required to resolve the correct 4-species tree when the proportion of gene trees matching the species tree (P, in blue) is low (40%). Prior 1 is (1,200) and Prior 2 is (1,1000). C) The number of gene required to resolve the correct 8-species tree when the proportion of gene trees matching the species tree (P, in blue) is low (10%). Prior 1 is (1,200) and Prior 2 is (1,1000). D) The increased confidence in the inferred 8-species tree when P increases.

3.3.2 The probability of gene trees matching the species vs. the posterior probability of the true species tree.

For a given sampling effort (e.g., 10 genes), the chance of correctly reconstructing the species tree increased as the proportion of gene trees matching the species tree increased (Figure 3.1D).

3.3.3 The comparison of our method with the concatenation method and the consensus method.

To compare our method with the other two methods, we used a new species tree of eight species to simulate sequence data. The true species tree was:
(sh , (sg , (sf , (se , (sd , (sc , (sa, sb):0.038095 #0.005427) :0.063516 #0.008777):0.069509 #0.039239) :0.073728 #0.047876):0.080717 #0.099646):0.128138 #0.107017):0.572334 #0.147364;

Thirty gene trees were generated from the above species tree. We simulated the DNA sequence of 500bp for each gene tree using the Jukes-Cantor substitution model. The DNA sequences were used to estimate the species tree for the concatenation method and our method.

The concatenation method yielded the following posterior distribution of the species tree:

$$tree_1[p = 0.982] = (sh, (sg, ((se, sf), (sd, (sc, (sa, sb))))));$$

$$tree_2[p = 0.009] = (sh, (sg, (sf, (se, (sd, (sc, (sa, sb)))))));$$

$$tree_3[p = 0.006] = (sh, (sg, (sf, (sd, (se, (sc, (sa, sb)))))));$$

$$tree_4[p = 0.001] = (sh, (sg, (se, (sf, (sd, (sc, (sa, sb)))))));$$

$$tree_5[p = 0.001] = (sh, (sf, ((sb, sd), (sc, (sg, (sa, se))))));$$

$$tree_6[p = 0.001] = (sh, ((sf, sg), (sd, (se, (sc, (sa, sb))))));$$

where p gives the posterior probability for the topology.

The Bayesian estimate of the species tree for the concatenation method is the first tree: (sh,(sg,((se,sf),(sd,(sc,(sa,sb))))));

The species tree as estimated by our method is:

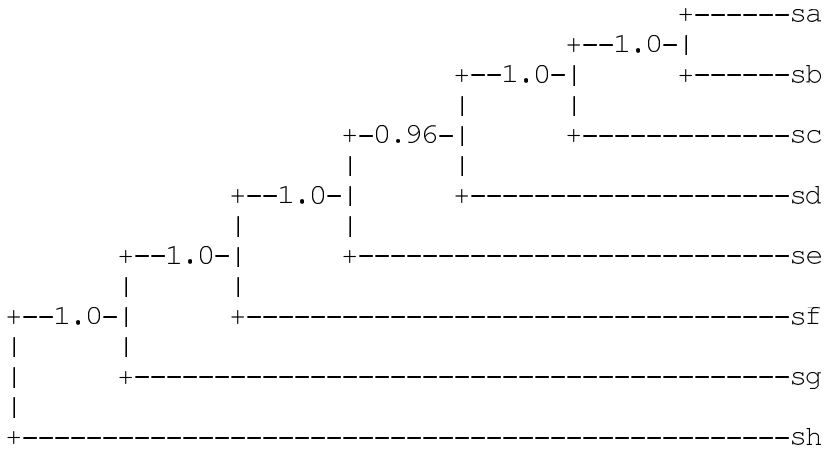

Figure 3.2: The estimate of the species tree using our method.

Our method estimated the true species tree with posterior probability 0.96, while the concatenation method estimated a different tree with posterior probability 0.98. The estimation of the species tree has two levels—the gene tree estimation and the species tree estimation. The gene tree estimation may be improved by increasing the sequence length, but the species tree estimation depends on the gene trees only. If the gene trees are in the anomaly zone [17] where the highest proportion of the gene trees do not match the true species tree, we suspect that the concatenation method is more likely to be biased and to estimate a wrong species tree. The pre-specified true species tree was intentionally made up to have short branch lengths and relatively large effective population sizes. According to the coalescent theory, this species tree is more likely to produce gene trees that are incongruent with the species tree, indicating the simulated gene trees may be in the anomaly zone. As a result, the concatenation method failed to recover the true species tree for this simulated data.

3.3.4 The number of genes vs. the estimates of population sizes and divergence times.

We simulated 3, 5, 10, 20, 40 and 80 gene trees from the true species tree (((si, sj) :0.0052 #0.001, ((sh , sg) :0.0068 #0.001 , (sf , se) :0.003 #0.001) :0.007 #0.001) :0.0092 #0.001 , (sd , (sc , (sa, sb):0.005 #0.001) :0.0072 #0.001):0.008 #0.001):0.018 #0.001; following the coalescent theory. These gene trees were then used to estimate effective population sizes θ and ancestral divergence times τ using exponential distribution with mean 0.005 and variance 0.000025 as the prior of θ. Repeat this procedure 100 times and calculate the averages and standard errors of the estimates of θ and τ for different number of gene trees.

population	θ	3 genes	5 genes	10 genes
1	0.001	0.002103(0.000444)	0.001758(0.000511)	0.001366(0.000380)
2	0.001	0.001683(0.000483)	0.001450(0.000432)	0.001197(0.000294)
3	0.001	0.002504(0.000574)	0.002129(0.000722)	0.001773(0.000677)
4	0.001	0.001759(0.000473)	0.001583(0.000503)	0.001308(0.000373)
5	0.001	0.001892(0.000475)	0.001616(0.000429)	0.001235(0.000323)
6	0.001	0.001888(0.000464)	0.001635(0.000420)	0.001336(0.000352)
7	0.001	0.001548(0.000376)	0.001308(0.000371)	0.001153(0.000271)
8	0.001	0.002735(0.000610)	0.002508(0.000674)	0.002093(0.000674)
9	0.001	0.001760(0.000350)	0.001600(0.000449)	0.001305(0.000290)
population	θ	20 genes	40 genes	80 genes
1	0.001	0.001201(0.000266)	0.001071(0.000174)	0.001023(0.000110)
2	0.001	0.001059(0.000195)	0.001030(0.000134)	0.001021(0.000091)
3	0.001	0.001387(0.000411)	0.001147(0.000216)	0.001065(0.000149)
4	0.001	0.001161(0.000224)	0.001098(0.000159)	0.001052(0.000092)
5	0.001	0.001125(0.000231)	0.001077(0.000149)	0.001038(0.000120)
6	0.001	0.001140(0.000235)	0.001076(0.000158)	0.001057(0.000108)
7	0.001	0.001065(0.000164)	0.001028(0.000115)	0.001023(0.000086)
8	0.001	0.001666(0.000717)	0.001307(0.000416)	0.001136(0.000222)
9	0.001	0.001134(0.000261)	0.001050(0.000171)	0.001020(0.000104)

Table 3.1: The Bayesian estimates of population sizes for the simulation data. The values in the parenthesis are the standard errors. The values in front of the parenthesis are the estimates of θ.

As expected, the results show that as the number of genes increases, the estimates of the population sizes and divergence times become closer to the true value and the standard error become smaller (Table 3.1 and Table 3.2). Overall, three genes appears to be adequate to deliver a relatively accurate estimate of the divergence time for ten species in this simulation model, but it requires at least 10 genes in order to have a good estimate of the population size. The effective population size estimates appear to be biased (Table 3.1), which might be caused by the prior we used.

population	τ	3 genes	5 genes	10 genes
1	0.018	0.017897(0.000210)	0.017961(0.000181)	0.017985(0.000071)
2	0.008	0.007989(0.000117)	0.007997(0.000070)	0.007998(0.000036)
3	0.0076	0.007555(0.000116)	0.007565(0.000088)	0.007571(0.000043)
4	0.005	0.005027(0.000167)	0.005008(0.000097)	0.004988(0.000039)
5	0.009	0.008969(0.000148)	0.008977(0.000095)	0.008999(0.000046)
6	0.0052	0.005217(0.000132)	0.005188(0.000079)	0.005192(0.000045)
7	0.007	0.006997(0.000108)	0.006985(0.000062)	0.006990(0.000034)
8	0.003	0.003088(0.000159)	0.003042(0.000118)	0.002996(0.000031)
9	0.0068	0.006705(0.000063)	0.006728(0.000048)	0.006763(0.000037)
population	τ	20 genes	40 genes	80 genes
1	0.018	0.017991(0.000020)	0.017997(0.000009)	0.018000(0.000006)
2	0.008	0.008000(0.000019)	0.008002(0.000011)	0.008001(0.000006)
3	0.0076	0.007591(0.000021)	0.007599(0.000013)	0.007599(0.000007)
4	0.005	0.005000(0.000025)	0.004999(0.000013)	0.005000(0.000006)
5	0.009	0.009000(0.000018)	0.009002(0.000012)	0.009001(0.000006)
6	0.0052	0.005197(0.000023)	0.005200(0.000013)	0.005200(0.000006)
7	0.007	0.006999(0.000014)	0.007000(0.000007)	0.006999(0.000003)
8	0.003	0.003003(0.000020)	0.003002(0.000015)	0.003000(0.000006)
9	0.0068	0.006780(0.000020)	0.006795(0.000011)	0.006801(0.000009)

Table 3.2: The Bayesian estimates of divergence times for the simulation data. τ is the true divergence time for each population. The values in the parenthesis are the standard errors. The values in front of the parenthesis are the averages of the estimates across the simulations.

CHAPTER 4

APPLICATIONS

We apply our method to three DNA sequence datasets. The finch dataset contains 4 species and 30 genes, a small number of species with a moderate number of genes. It is then easy to present and compare the posterior distributions of the species tree and gene trees for the finch dataset since there are only 3 possible topologies relating the 4 species.

The yeast dataset has a small number of species (8) with a large number of genes (106), which indicates that the dataset contains strong information about the species tree. As a consequence, we expect that the estimate of the species tree for the yeast dataset will have a strong posterior probability support.

The macaque dataset includes a moderate number of species (19) with a small number of genes (4). The dataset is the mixture of different types of genes – autosomal genes, a mitochondria gene and a Y-chromosomal gene. Y-chromosomal and mitochondrial genes are uniparentally inherited and haploid making their effective population sizes one-fourth that of autosomal markers. To analyze the macaque dataset, our method has been extended to allow different effective population sizes for different types of genes.

4.1 Finch data analysis.

We first apply the new method to a multilocus nucleotide dataset from birds recently published by Jennings and Edwards (2005) [58]. They obtained the allelic data of 30 loci (Pa-0, $\cdots$, Pa-30) from one individual per population of *P.acuticauda* (species 1), *P.hecki* (species 2) and *P.cincta* (species 3). They also included sequences from a more distant relative; the zebra finch (*P.guttata*; species 4) as outgroup. A total of 30 anonymous loci were developed ranging in size from 216 to 825 bp. They performed a four-gamete test [51] and the result showed that the overall incidence of intralocus recombination in the data appears low which supports one of the assumptions in our model that there is no intralocus recombination. Jennings and Edwards also used the assumed species tree topology previously supported by morphological and mtDNA studies and employed a multilocus coalescent approach to infer the effective population sizes and divergence times.

4.1.1 Data analysis

(1) Posterior distributions of gene tees for 30 genes using the independent prior.

The posterior distributions of gene trees are estimated in MrBayes assuming independent loci. HKY85 [49] was selected as the substitution model that best fit the data according to a hierarchical likelihood ratio test [33]. Since the position of the species 4 is fixed as the outgroup, there are only three possible topologies, (2,(1,3)), (3,(1,2)), (1,(2,3)). From Table 4.1, there are 15 genes out of 30 whose estimates of the gene tree support the tree (3,(1,2)). The average probability for (3,(1,2)) across all 30 genes is 0.434. The corresponding probabilities for the other two possible trees are 0.205 and 0.361. The tree (3,(1,2)) has more support from the gene trees than

(2,(1,3)) and (1,(3,2)). One way to estimate a species tree from multiple gene trees when there are three taxa is via a majority-rule criterion, whereby the gene tree whose topology is found most frequently is presumed to reflect the topology of the species tree. The majority-rule estimate of the species tree is thus the second tree, (3,(1,2)).

(2) Estimation of the topology of the species tree using the concatenation method.

In this case the multilocus sequences are concatenated into a single sequence. The concatenated data was analyzed in MrBayes with a HKY85 substitution model. The prior for the topology was taken to be a uniform distribution, and branch lengths are assumed to be independently distributed exponentials. The estimated topology of the species tree is in the Table 4.1. It matches the majority-rule tree in (1) and its posterior probability is essentially one. Since there are a total of 16119 bp in the concatenated data, it is not surprising that the estimated tree was converged to be so highly resolved.

(3) Bayesian estimation of gene tees, topology of species trees, effective population sizes and divergence times using the proposed method.

The finch dataset was analyzed in MrBayes. The posterior distribution of gene trees was estimated with a HKY85 substitution model and the joint prior of gene trees across 30 genes. The estimated joint gene trees were then used to reconstruct the species tree using MCMC as implemented in the program Bayesian Estimation of Species Tree (BEST). Three different priors of effective population sizes were used to evaluate the effect of the priors on the posterior distributions. The priors are exponential distributions with means 1, 0.1, 0.0072, and 0.00072. The medians of these four priors for effective population sizes (in the units of substitutions per site) are then 0.693, 0.069, 0.005, and 0.0005. They reflect the user's initial guess about the

	Independent prior			Joint prior		
	(2,(1,3))	(3,(1,2))	(1,(2,3))	(2,(1,3))	(3,(1,2))	(1,(2,3))
Pa-1	0.184	0.671±0.001	0.146	0.171	0.683±0.025	0.146
Pa-2	0.337	0.353±0.002	0.309	0.299	0.375±0.026	0.326
Pa-3	0.062	0.88±0.001	0.058	0.056	0.915±0.015	0.029
Pa-4	0.331	0.331±0.001	0.337	0.221	0.452±0.027	0.327
Pa-5	0.319	0.319±0.001	0.361	0.264	0.398±0.0264	0.338
Pa-6	0.012	0.966±0.001	0.022	0.047	0.894±0.0166	0.059
Pa-7	0	1±0	0	0	1±0	0
Pa-8	0	1±0	0	0	1±0	0
Pa-9	0.042	0.912±0.001	0.046	0.026	0.935±0.0133	0.038
Pa-10	0.222	0.547±0.002	0.232	0.117	0.699±0.0248	0.184
Pa-11	0	1±0	0	0	1±0	0
Pa-12	0.319	0.353±0.002	0.327	0.293	0.449±0.0269	0.258
Pa-13	0.493	0.503±0.002	0.004	0.257	0.743±0.0236	0
Pa-14	0.242	0.503±0.002	0.255	0.254	0.497±0.027	0.249
Pa-15	0.325	0.349±0.002	0.325	0.151	0.578±0.0196	0.271
Pa-16	0.335	0.333±0.001	0.331	0.233	0.496±0.0270	0.271
Pa-17	0.042	0.02±0.000	0.938	0.073	0.156±0.0196	0.772
Pa-18	0	0±0	1	0	0±0	1
Pa-19	0	0±0	1	0	0±0	1
Pa-20	0	0.002±0.000	0.998	0	0±0	1
Pa-21	0	0±0	1	0	0±0	1
Pa-22	0.04	0.076±0.001	0.884	0.045	0.085±0.0151	0.87
Pa-23	0.014	0.064±0.001	0.922	0.019	0.046±0.0113	0.935
Pa-24	0	1±0	0	0.002	0.998±0.0024	0
Pa-25	0.311	0.339±0.001	0.349	0.232	0.503±0.027	0.265
Pa-26	0.782	0.212±0.001	0.006	0.482	0.5±0.027	0.018
Pa-27	1	0±0	0	1	0±0	0
Pa-28	0.389	0.305±0.001	0.305	0.298	0.431±0.0267	0.271
Pa-29	0.01	0.653±0.002	0.337	0.001	0.739±0.0237	0.26
Pa-30	0.333	0.327±0.001	0.339	0.164	0.68±0.0252	0.156
Average	0.205	0.434	0.361	0.157	0.508	0.335
concatenation	0	1±0	0			
gamma(1,139)	0.08	0.88±0.018	0.04			
gamma(1,1389)	0.03	0.95±0.007	0.02			
gamma(1,10)	0.08	0.89±0.019	0.03			
gamma(1,1)	0.01	0.94±0.015	0.05			

Table 4.1: The posterior distributions of gene trees and species trees in the finch data set. The estimated posterior probability ± sd for the topology (3,(1,2)) is listed for each gene. The row labeled concatenation is the posterior probabilities of species trees for the concatenation method. The last several rows give the posterior probabilities from the BEST method with different priors for θ: gamma(1,139), gamma(1,1389), gamma(1,10) and gamma(1,1).

population sizes. To convert the posterior medians of these parameters to estimates of the posterior speciation times in years and effective population sizes, we assume the mutation rate is $3.6 * 10^{-9}$ as in Jennings and Edwards (2005) [58].

The same species tree with strong support is estimated no matter which of the priors we use (Table 4.1). Our estimate of the species tree agrees with the one estimated by the concatenation method except that the support for the clade (1,2) is essentially one for the concatenation method and the support of the same clade is approximately 0.88 for the BEST method (Table 4.1).

To compare the BEST method with the concatenation method, we estimate the Bayes factor using the harmonic mean of the likelihood. The logarithm of the Bayes factor is estimated to be about 300. Although the harmonic mean method can be somewhat unstable and sensitive to the lowest value of the likelihood, it works here since the log likelihoods for the two different methods are well separated (Figure 4.1). The Bayes factor suggests that the coalescent model fits the data better than the concatenation method.

The posterior estimates of the divergence times are similar for the different priors (Table 4.2), indicating a strong signal in the data for these parameters. On the other hand, the posterior distribution of the population sizes does appear to be sensitive to the prior chosen. The estimate is strongly correlated to the median of the prior for the clade (1,2), whereas the clade (1,2,3) is relatively insensitive to the priors (Table 4.2).

The gene trees are correlated as a consequence of their joint dependence on the species tree. We should use a joint distribution to formulate the prior of gene trees from different genes. In our model, the joint distribution is derived from coalescent

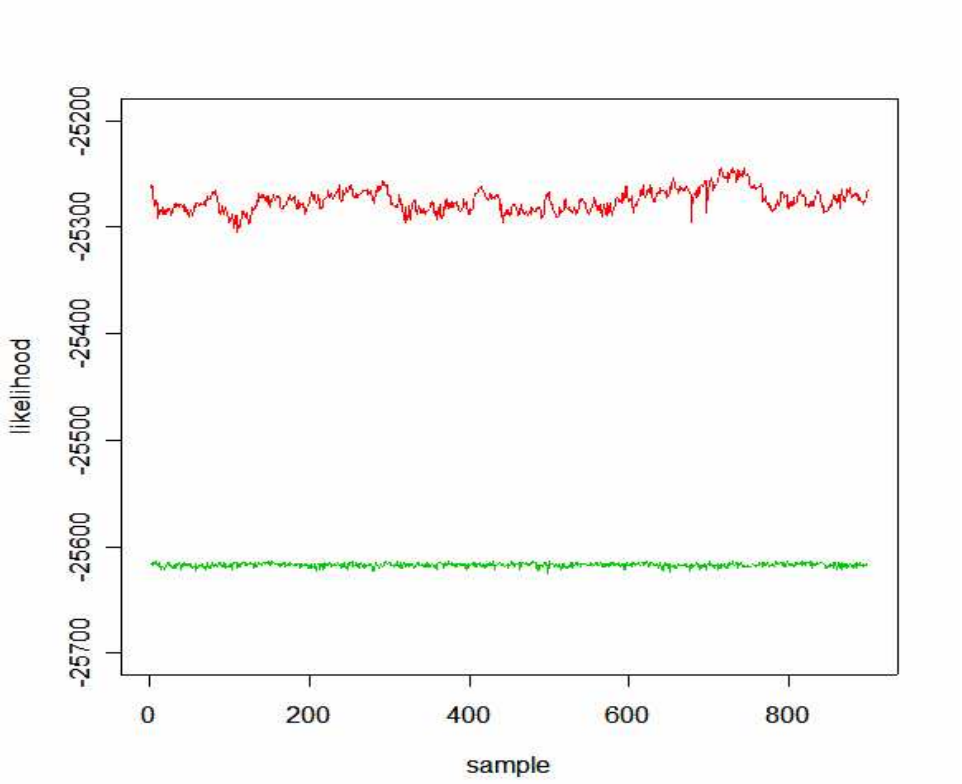

Figure 4.1: The log likelihood curves of two analyses for the finch data set. The pink curve is the log likelihood for our model. The blue curve is the log likelihood for the concatenation method. They are well separated with the proposed model having much greater likelihood than the concatenation method.

theory. Let G_i be the gene tree for gene i and S be the species trees. The joint distribution of gene trees is given by

$$f(G) = f(G_1 \cdots f_k) = \int_S f(G_1|S,\theta) \cdots f(G_k|S,\theta) dS d\theta$$

.

where $f(G_i|S,\theta)$ follows coalescent theory. $f(G_1 \cdots G_k)$ tends to put more weight on gene trees with similar topologies and branch lengths. This can be seen from the posterior probabilities of the gene trees in Table 4.1. There are 22 genes supporting the tree (3,(,1,2)) under the joint prior, whereas only 15 genes in Table 4.1 support that tree when the independent prior is used. The average support probability for (3,(2,1))

is 0.508 which increases by 0.074 from the average support for the independent prior. Interestingly, we can see the pattern of the change of the posterior probability of gene trees. Consider genes Pa-4, Pa-5 and Pa-18. For genes Pa-4 and Pa-5, both posteriors under the independence model favor the third topology in Table 4.1. However, after adjusting for the species coalescent process, their posteriors change to favor the second topology, which is the Bayesian estimate for the majority of genes. This is because gene trees are correlated and the topology of a particular gene tree depends on the gene trees for other genes. If a majority of genes support the same topology, it will make the rest of the genes more likely to have a similar topology. But if the support is too strong as for the gene Pa-18 that strongly supports the third topology with probability near 1, its posterior may not be dramatically changed even if the joint prior is used. A similar pattern is seen with Pa-21, 16 and 30 in Table 4.1.

Our estimate of the species tree agrees with the assumed species phylogeny found by Jennings and Edwards (2005) [58]. The posterior probability of the species tree is around 0.9 no matter what prior we used. The estimate of the divergence time for (1,2) using our method is similar to the estimate given by Jennings and Edwards. However, our estimate for the clade ((1,2),3) is 0.00418 which is higher than Jenning and Edwards' estimate (0.00254). For the population size, both methods have the estimate for the clade ((1,2),3) around 0.005. However, the estimates of the population size of clade (1,2) are sensitive to the prior for both techniques.

Exponential mean 0.00072	Divergence times	Population sizes
(1,2)	0.00408(0.00277, 0.00457)	0.00218(0.00072, 0.00553)
(1,2,3)	0.00449(0.00344, 0.00547)	0.00407(0.00252, 0.00604)
Exponential mean 0.0072	Divergence times	Population sizes
(1,2)	0.00297(0.00147,0.00389)	0.00693(0.00069,0.02813)
(1,2,3)	0.00418(0.00325,0.00523)	0.00506(0.00290, 0.00837)
Exponential mean 0.1	Divergence times	Population sizes
(1,2)	0.00235(0.00100,0.00376)	0.0155 (0.00149,0.23625)
(1,2,3)	0.00418(0.00326,0.00493)	0.00481(0.00292, 0.00783)
Exponential mean 1	Divergence times	Population sizes
(1,2)	0.00249(0.00065,0.00262)	0.01237 (0.00310,0.44104)
(1,2,3)	0.00429(0.00343,0.00525)	0.00503(0.00309, 0.00799)

Table 4.2: Estimates of the ancestral population sizes and divergence times for different priors in the finch data set. The priors for θ are Exponential with means 1/1389, 1/139, ,1/10, or 1. For each prior, the estimates of the divergence times and population sizes of a particular ancestral population are listed along with a 95% credible region. (1,2) represents the ancestral population of species1 and species2. (1,2,3) is the ancestral population of species1, species2 and species3.

4.2 Macaques Data Analysis.

Tosi and Morales (2003) [122] isolated total genomic DNA of 63 macaques from 19 species and eight outgroup taxa. The DNA sequences were obtained from Y-Chromosomal loci, mtDNA, C4 long Intron 9 and IRBP Intron 3. In their analysis, the ML tree was estimated for each gene assuming an HKY85+G substitution model. The four different gene trees were used to make inference on the pattern of the species tree, but no method was available to combine the data, and concatenating the sequences was deemed inappropriate. Divergence times were estimated only for the Y-Chromosomal and mitochondrial trees.

Here we analyze a reduced data set of one randomly chosen allele from each of the 19 species, including one outgroup taxa (*T. Gelada*), for which there is data on all four "genes". The modified data was analyzed using the proposed method to estimate the posterior of gene trees and species trees. Further, a sensitivity analysis was performed to investigate the influence of each gene on the overall estimate of species trees.

4.2.1 Material and methods.

(1) Estimate of the species tree using the BEST algorithm.

The effective population sizes are parameters in the likelihood of gene trees given species trees. In theory, Y-chromosomal and mitochondrial genes are uniparentally inherited and haploid making their effective population sizes one-fourth that of autosomal markers. This dataset is a mixture of Y-chromosomal, mitochondrial genes and autosomal genes. According to the coalescent theory, the probability that the gene tree matches the species tree depends on the ratio of branch lengths and the effective population size. Thus, to make the data from the four genes comparable, the 1-to-4 effective population size adjustment based on the mode of inheritance was made in our analysis. The concatenation method does not apply to this example because the genes in the data have different effective population sizes, and a different mode of inheritance [86] [87] [108].

Fooden defined four species groups for macaques according to distinct forms of male reproductive anatomy [41]. The species groups include the silenus group, fascicularis group, sinica group and arctoides group. Our estimate of the species tree identified the silenus group with a moderate posterior probability (Figure 4.2). The

	independent-species	joint-species
Y-Chromosome	0.781 ± 0.089	0.656 ± 0.085
mtDNA	0.739 ± 0.092	0.646 ± 0.084
C4 Intron 9	0.779 ± 0.054	0.659 ± 0.078
IRBP Intron 3	0.838 ± 0.052	0.659 ± 0.070

Table 4.3: The average ± st.dev. of distances between two posterior distributions for each gene in the macaques data set. There are three posterior distributions for each gene, the posterior of species trees, and the posterior of gene trees with the independent gene model and the posterior of gene trees with the joint coalescent-based model. The average distances between the posterior of species trees and the posterior with the independent prior (denoting by independent-species) as well as the posterior of species trees and the posterior with the joint prior (denoting by joint-species) are calculated by Phylip (Felsenstein, 2003) [36] using the symmetric distance measure [103].

species in the other three groups are not well resolved. This indicates either that Foodens clades are poorly defined or inadequate information in the dataset and that more genes or alleles may be needed for estimating the species tree of macaques.

It is interesting to compare the posterior of the gene trees with the posterior of species trees. Let us define a measurement of the distance between two random trees, D(T1,T2), to be the symmetric distance between the tree T1 and tree T2 [103]. This simply measures the number of branches the two trees do not have in common. Table 4.3 provides the average ± standard deviation of the distribution of distances between the gene tree, T1, and the species tree, T2, based on the posterior distributions computed under both the independence model and the coalescent model. The distances for the independent gene model are larger than the distances for the joint coalescent-based model for all the four genes. As expected, the result suggests that

the joint model makes the gene trees closer to the species tree than the independent gene model.

(2) Comparison of coalescent-based model with the independent gene model. Our method assumes the joint estimate of the posterior of gene trees must be compatible with the species tree while common analyses assume that loci are independent. To evaluate the effect of different priors on the posterior distribution of gene trees, we want to know how different the posterior distributions of gene trees will be using the two different priors. To examine this issue, we introduce a theorem by Maa (1996) [76].

We want to examine the difference between two high dimensional distributions F1 and F2 which here are the two distributions to be examined. Let X1 and X2 be independent and identically distributed random quantities from F1 and independent of Y1 and Y2 from F2. Take D(.,.) to be any appropriately chosen distance function. The theorem posits that F1=F2 if and only if the one-dimensional quantities D(X1 , X2) = D(Y1 , Y2)= D(X1 , Y1) in distribution.

To apply the idea of the theorem, we calculated the three distances (two within group distances and the between groups distance) for each gene in Table 4.4. The smaller joint-joint distances indicate less variability in the trees from the joint prior compared with the independent prior. The larger independent-joint distances indicate the high degree of separation of the two posterior distributions of gene trees for this data set. The results show that the three distances are quite different for all genes, indicating that the joint prior and the independent prior result in two significantly different posterior distributions of gene trees for this dataset.

(3) Sensitivity analysis.

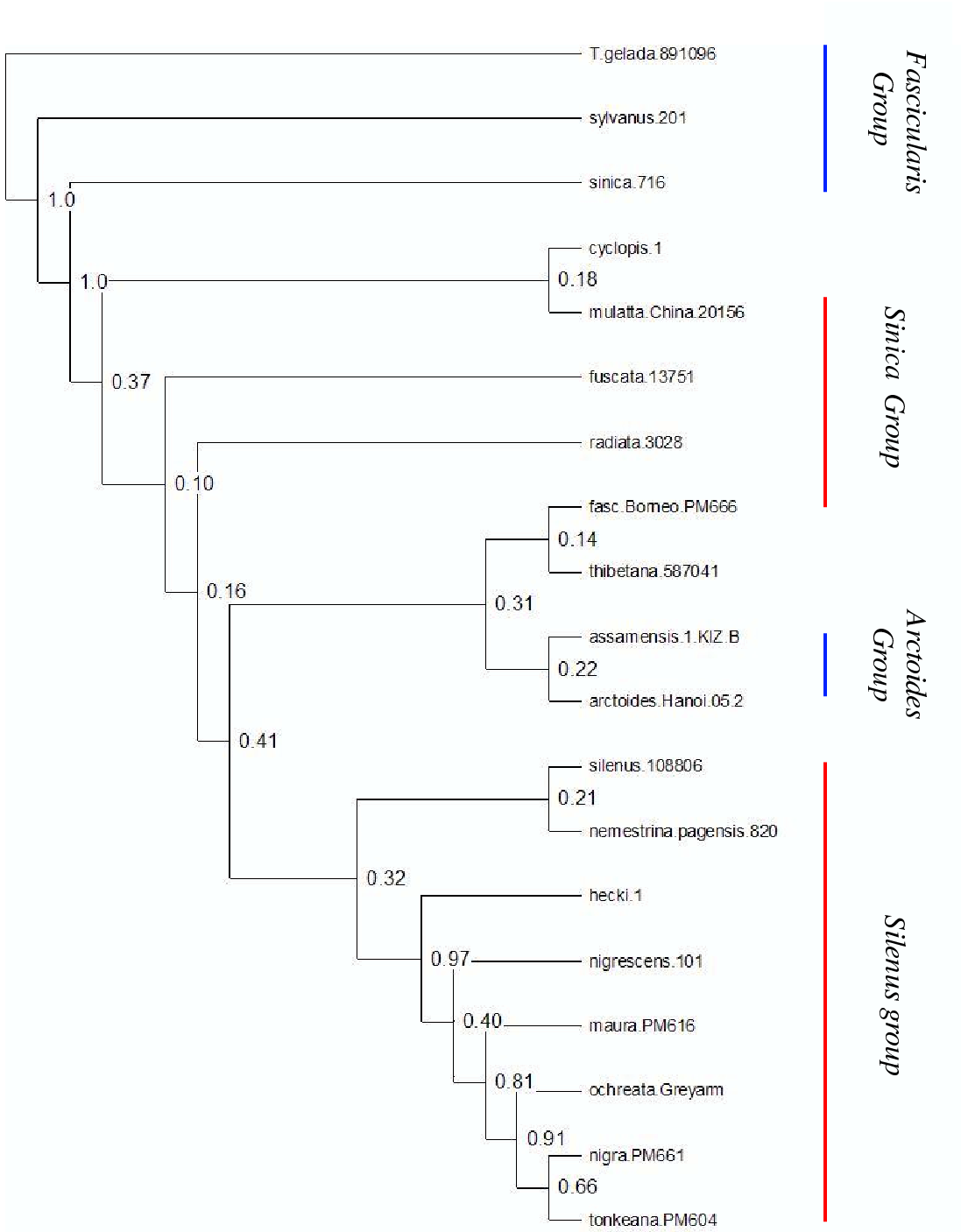

Figure 4.2: The estimate of species tree for macaques using the BEST method. This is the consensus tree of the sample trees from the posterior distribution of species trees. *T.gelda* is the outgroup. The species tree has 4 species groups.

	independent-independent	joint-joint	independent-joint
Y-Chromosome	0.309 ± 0.067	0.164 ± 0.066	0.356 ± 0.070
mtDNA	0.215 ± 0.077	0.070 ± 0.062	0.272 ± 0.087
C4 Intron 9	0.319 ± 0.067	0.237 ± 0.062	0.501 ±0.047
IRBP Intron 3	0.257 ± 0.068	0.101 ± 0.068	0.362 ±0.052

Table 4.4: The average ± st. dev. of distances between two posterior distributions for each gene in the macaques data set. There are two posterior distributions for each gene, the posterior of gene trees assuming independent genes and the posterior of gene trees with the joint coalescent-based model. The average distances between each posterior and itself (denoted by independent-independent or joint-joint) are calculated in Phylip. The average distance between the two different posterior (denoted by independent-joint) is also calculated in Phylip.

S1	S2	S3	S4
0.683 (0.597,0.768)	0.682 (0.607, 0.757)	0.664 (0.583, 0.746)	0.663 (0.580, 0.747)

Table 4.5: The average and 95% credible regions for distances between the posterior distribution of species trees estimated with all 4 genes and the posterior distribution of species trees estimated by leaving one gene out in the macaques data set.

Genes may have different influence on the posterior distribution of species trees. We examined the potential gene-by-gene sensitivity of our results by eliminating each single gene from the analysis in turn and re-estimating the posterior of species trees. Let S1, S2, S3, and S4 be the posterior of species trees estimated without the Y-Chromosome, mtDNA, C4-Intron 9, and IRBP Intron 3 data respectively. Table 5 displays the distances between each Si and S (the posterior of species trees using all 4 genes). The mean distances for all four genes are comparable and the credible intervals for the four distances almost entirely overlap. This suggests that the estimation of the species tree is not overly subject to the strong influence of any single gene in this dataset.

4.3 Yeast data analysis.

Antonis Rokas kindly supplied the full 106-gene data set published in the original paper [105]. The dataset includes 8 species of yeast: *S.cerevisiao, S.paradoxus, S.mikatae, S.kudriavzevii, Sbayanus, Skluyven, S.castellii* and *C.albicans*. The *C.albicans* is the outgroup.

MrBayes was run for 80 million cycles per analysis to estimate the posterior distributions of gene trees. A $GTR + \Gamma + I$ model [118] was used for all analyses. We used Exponential distribution with mean 0.005 as the prior for θ: We also tried a prior of Exponential(10) and Exponential(1000) for θ.

The molecular clock assumption can be relaxed if we can find a way to convert a no-clock tree to a clock tree. The subroutine DNAMLK in Phylip is able to force a pre-specified no-clock tree to have a clock by changing the branch lengths to be ultrametric. We recoded DNAMLK and added it into MrBayes. We allowed MrBayes

to propose a no-clock tree and use it to calculate the likelihood $f(D|G, \lambda)$. The no-clock tree was rooted by the outgroup and transformed to a clock tree using the DNAMLK function we added in MrBayes. The clock tree was then used to calculate the joint prior $f(G)$. However, this technique of converting a non-clock tree to a clock tree is fairly arbitrary and lacking biological meaning. A better way to relax the molecular clock might be to add the mutation rates along lineages as parameters into the model and allow the lineages to have different mutation rates. We will discuss this issue in the next chapter.

We used four different models - independent model with molecular clock, independent model without molecular clock, joint model with molecular clock and joint model without molecular clock, to estimate the posterior distribution of the species tree and gene trees for the yeast dataset. The independent model assumes that the 106 genes are independent. On the contrary, the joint model (our model) assumes that the genes are correlated through the species tree.

4.3.1 Posterior distributions of gene trees

The posterior distributions of gene trees for each of the 106 genes in the yeast data set is presented in Figure 4.3 for each of the four models: 1) independent model with a molecular clock, 2) joint model with a molecular clock, 3) independent model without a clock, 4) joint model without a clock. The trees in the Figure 4.3 are designated as 1-24 as follows:

1. (8,(7,(6,(5,(4,(3,(1,2)))))));
2. (8,(7,(6,((4,5),(3,(1,2))))));
3. (8,((6,7),(5,(4,(3,(1,2))))));

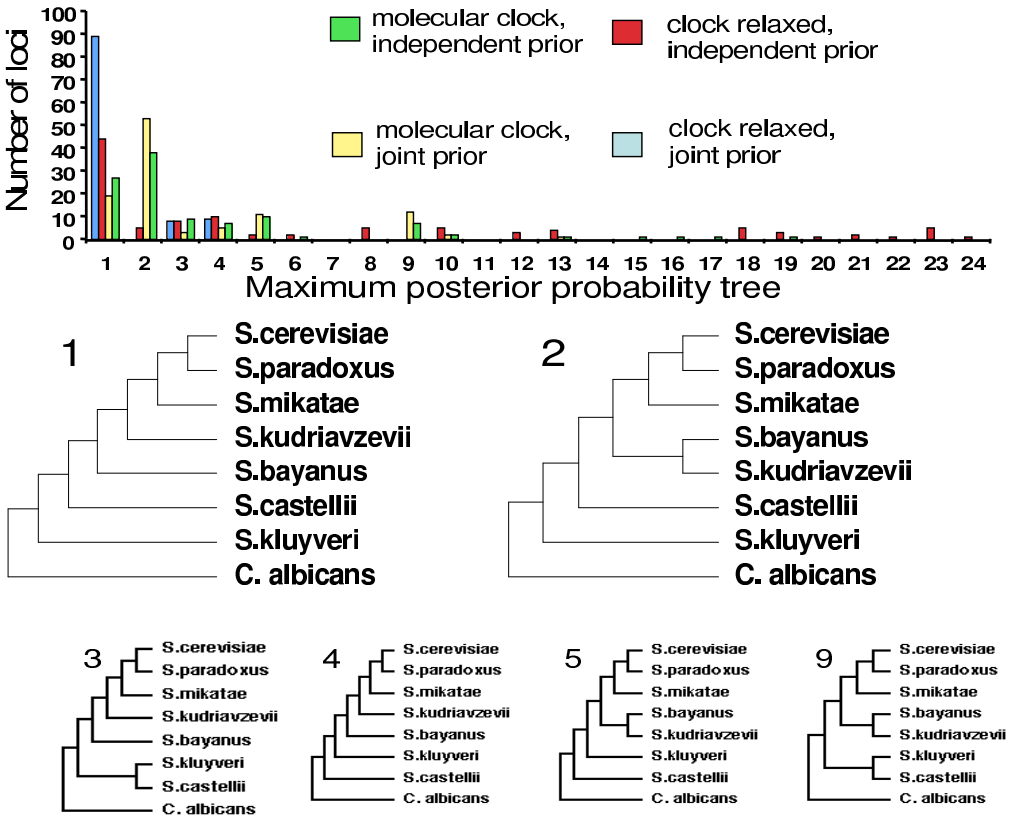

Figure 4.3: The distribution of gene trees for the 106-gene yeast data set. The number of genes (y-axis) yielding each of 24 topologies according to the maximum posterior probability criterion (x-axis) is shown for each of four analyses: independent (green) and joint model (yellow) with a molecular clock, and independent (red) and joint model (blue) without a molecular clock. The two most commonly encountered maximum posterior probability trees are shown below, with the next four most common shown in the bottom row.

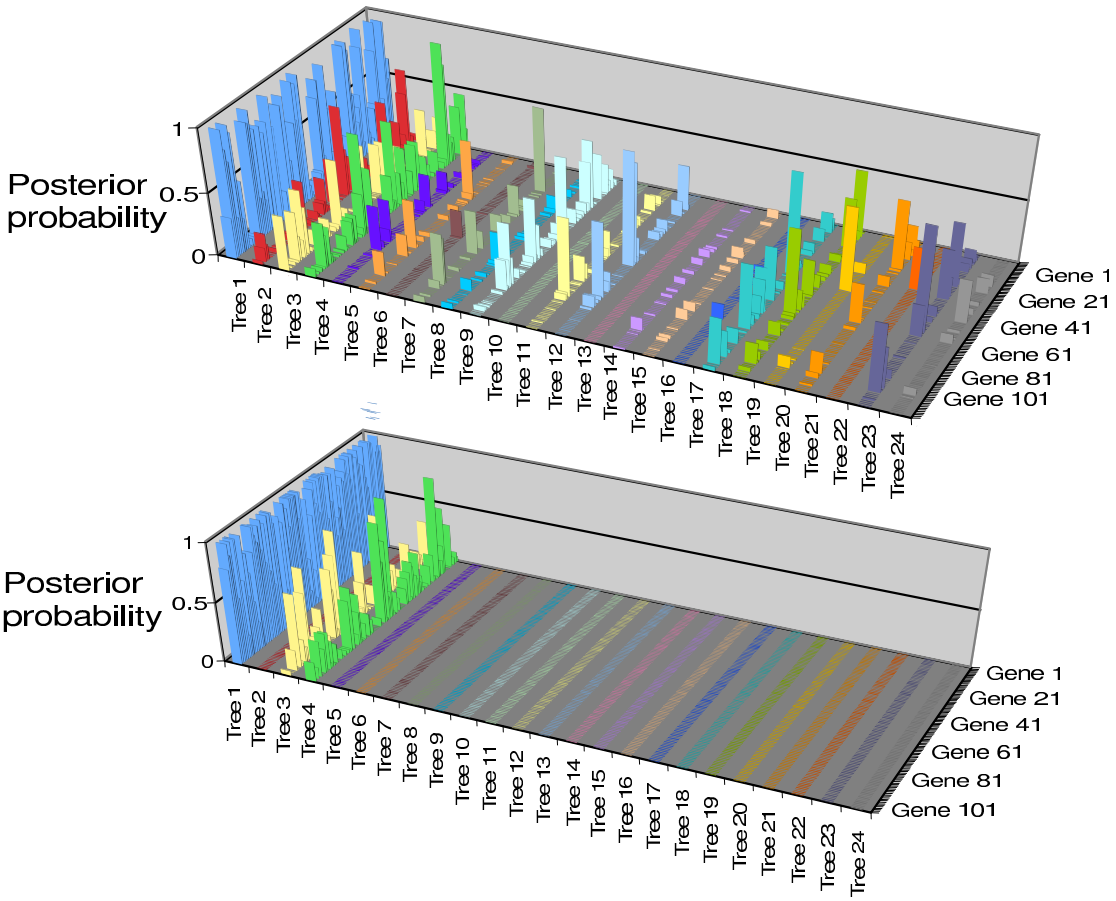

Figure 4.4: Shifting phylogenetic landscapes for gene trees under different models. The complete posterior probability distributions for the independent (top) and joint (bottom) model without a molecular clock are shown. Notice the small number of gene trees that receive substantial posterior probability in the joint model analysis as compared with the independent model.

4. (8,(6,(7,(5,(4,(3,(1,2)))))));
5. (8,(6,(7,((3,(1,2)),(4,5)))));
6. (8,(6,(7,(3,((1,2),(4,5))))));
7. (8,(7,(6,(5,((3,4),(1,2))))));
8. (8,((3,((4,5),(1,2))),(6,7)));
9. (8,((6,7),((3,(1,2)),(4,5))));
10. (8,(7,(6,(4,(5,(3,(1,2)))))));
11. (6,(8,(7,((4,5),(3,(1,2))))));
12. (8,((6,7),(4,(5,(3,(1,2))))));
13. (8,(6,(7,(4,(5,(3,(1,2)))))));
14. (6,(8,(7,(5,(4,(3,(1,2)))))));
15. (8,(((1,2),(3,(4,5))),(6,7)));
16. (8,(6,(7,((1,2),(3,(4,5))))));
17. (8,(7,(6,(5,(3,(4,(1,2)))))));
18. (8,(7,(6,(3,((1,2),(4,5))))));
19. (8,(7,(6,((3,(4,5)),(1,2)))));
20. (8,(3,((4,(5,(6,7))),(1,2))));
21. (8,(7,(6,(2,(1,(3,(4,5)))))));

22. (8,(7,(5,(4,(3,(2,(1,6)))))));

23. (8,(7,(6,(1,(2,(3,(4,5)))))));

24. (8,((6,7),(1,(2,(3,(4,5))))));

The previous study obtained 23 different topologies for the 106 genes when analyzed by parsimony or maximum likelihood. Using the Bayesian method, we found that consideration of 24 distinct topologies was sufficient to explain on average of more than 95% of the posterior distribution of gene trees for all analyses. The posterior distribution of gene trees under the independent model with a molecular clock was populated by 13 distinct topologies across all 106 gene (Figure 4.4). In this analysis, the highest probability tree for only 27 of 106 genes matched the concatenated tree published by Rokas (topology 1, Figure 4.3), whereas 38 genes yielded a maximum posterior probability gene tree in which *S. kudriavzevii and S. bayanus* form a clade (topology 2, Figure 4.3). As expected, the posterior distribution of gene trees was noticeably more concentrated on a few (eight) trees under a joint prior with a clock. We found that only 10 of the 106 genes in the data set were consistent with a molecular clock by a likelihood ratio test, suggesting that many of the gene trees estimated under a clock could be erroneous. We therefore developed a Markov Chain Monte Carlo approach for estimating gene trees without a molecular clock. The effect of the joint model in concentrating the probability distribution of gene trees around a few topologies is even more evident without the constraints of a clock (Figure 4.4 and Figure 4.4). Under these conditions, only three gene trees are plausible, many fewer than implied by the independent analyses. Moreover, under a joint model the highest probability tree for 89 of the 106 genes matched the concatenated tree of the 106

genes, and the next most common alternative was favored by only 8 genes (Figure 4.4 and Figure 4.3).

4.3.2 Bayes Factor analysis.

We used the harmonic means of the likelihoods to estimate Bayes factors between the four models and the concatenation model. The log Bayes factor favoring the joint model with relaxed clock over the next best model (independent model with a relaxed clock) was 130, which strongly supports the joint prior without molecular clock over the other models. The independent model was actually favored over the coalescent model when a molecular clock was enforced. The concatenation model was the worst model.

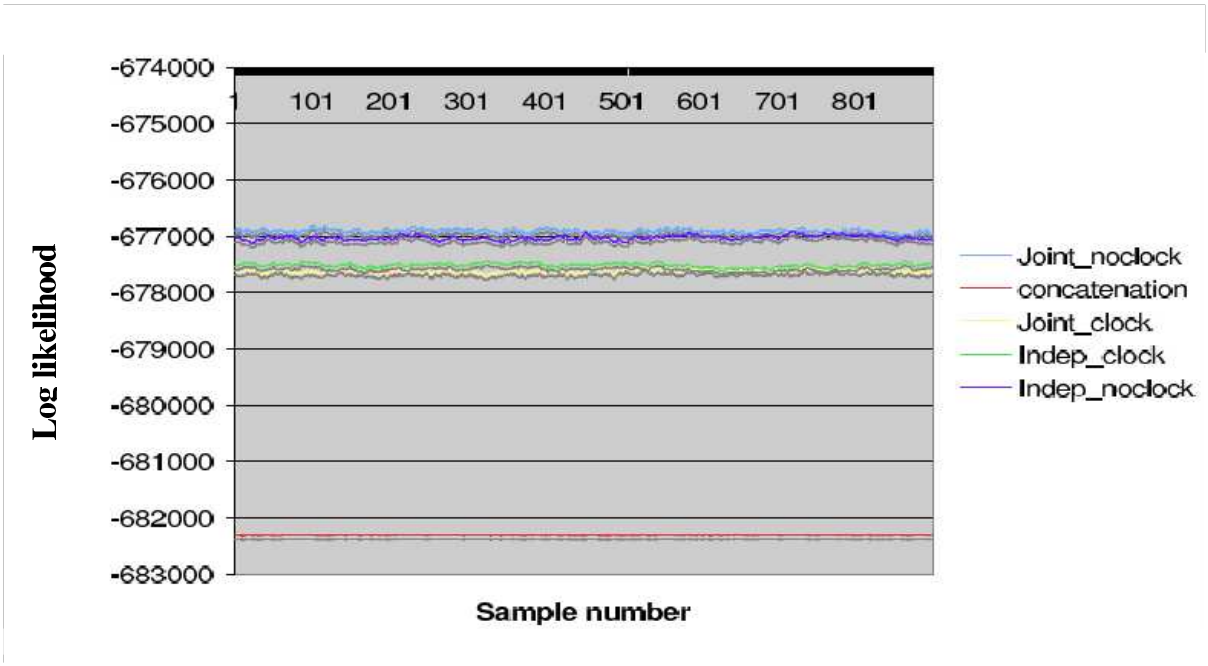

Figure 4.5: Comparison of likelihoods of five priors on the yeast data set

4.3.3 Estimated species tree.

For all analyses except the joint model without a molecular clock, the estimated species tree was tree 2 in Figure 4.3. By contrast, the species tree estimated with a joint model and a relaxed clock was tree 1 in Figure 4.3. In the case of the independent model and joint model with a clock, the species tree had posterior probability essentially 1 on all nodes. In the case of the independent model without a molecular clock, the majority rule consensus tree is displayed in Figure 4.6. The tree has at each node a number indicating how often the group which consists of the species descended from the node occurred in the estimated posterior distribution of the species tree.

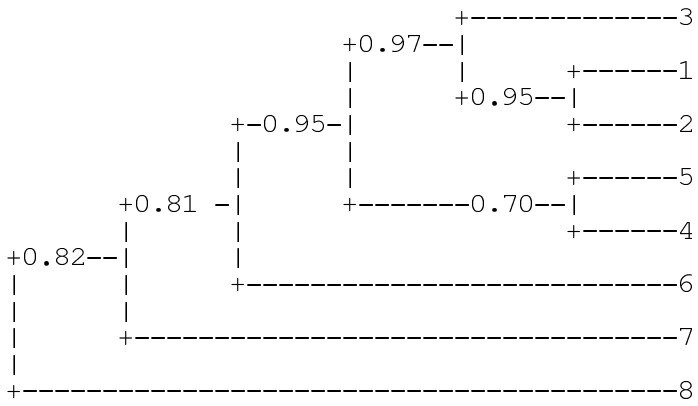

Figure 4.6: Estimate of species tree for the independent prior without a molecular clock.

For the joint prior without a molecular clock, the estimated species tree had the same topology but with a posterior probability of 1 on all nodes.

	θ(mean±sd)	divtime(mean±sd)
(1,2,3,4,5,6,7,8)	0.515(0.483,0.547)	0.581(0.600,0.562)
(1,2,3,4,5,6,7)	0.115(0.101,0.129)	0.327(0.319,0.335)
(1,2,3,4,5,6)	0.143(0.119,0.167)	0.235(0.224,0.246)
(1,2,3,4,5)	0.006(0.004,0.008)	0.110(0.108,0.112)
(1,2,3,4)	0.005(0.004,0.006)	0.088(0.086,0.090)
(1,2,3)	0.003(0.002,0.004)	0.063(0.062,0.064)
(1,2)	0.005(0.004,0.006)	0.038(0.037,0.039)

Table 4.6: Estimate of θ and divergence times for the joint prior without a molecular clock

	θ(mean±sd)	divtime(mean±sd)
(1,2,3,4,5,6,7,8)	0.459(0.432,0.487)	0.354(0.346,0.362)
(1,2,3,4,5,6,7)	0.120(0.106,0.134)	0.212(0.207,0.217)
(1,2,3,4,5,6)	0.126(0.106,0.146)	0.170(0.165,0.175)
(1,2,3,4,5)	0.015(0.013,0.017)	0.063(0.062,0.064)
(4,5)	0.024(0.017,0.031)	0.053(0.052,0.054)
(1,2,3)	0.002(0.001,0.003)	0.049(0.048,0.050)
(1,2)	0.003(0.002,0.004)	0.029(0.028,0.030)

Table 4.7: Estimate of θ and divergence times for the joint prior with a molecular clock

4.3.4 Estimates of θ and divergence times.

The estimates for θ and species divergence times are listed in Table 4.6.

Clearly the estimates of θ increase to unrealistically high values in progressively ancestral species. This may indicate a problem with the molecular clock assumption and the use of DNAMLK function to arbitrarily force a non-clock tree to a clock tree.

4.3.5 How many genes are required to estimate the species tree for yeasts by the Bayesian species tree method?

To determine the efficiency of the Bayesian species tree method, we first randomly chose 8 yeast genes and use the gene trees from their posterior distributions to build species trees. We then repeated this ten times to see how many samples out of ten can recover the species tree, each time re-estimating the posterior distribution of gene trees with a joint prior without a clock using 5,000,000 MCMC cycles. For each of these 10 replicates, we found that every estimated species tree was the same as topology 1 (Figure 4.3) with an average posterior probability of each node > 0.95. We then repeated this for 5 genes instead of 8. We found that in one of the 10 estimated species trees, the topology is different from topology 1 (Figure 4.3), although the other 9 correctly estimated topology 1 (Figure 4.3) with high support > 0.95. We therefore conservatively estimate that 8 genes in the yeast data set are sufficient to estimate the correct species tree with high probability.

CHAPTER 5

DISCUSSION AND FUTURE RESEARCH

The Bayesian hierarchical model we have employed adopts coalescent theory to formulate the distribution of gene trees given the species trees. Maddison and Knowles [81] have found by a simulation study that increasing the number of loci gives more accurate estimate of the species tree under the assumption that deep coalescence is the only reason for the conflicts between the gene tree and species tree. However, there are other biological factors that can affect the correspondence of species trees and gene trees. For example, horizontal transfer and gene duplication/loss may cause conflicts between gene trees and species trees. Unfortunately, it is challenging to model the underlying mechanism of horizontal transfer or gene duplication/loss without encountering problems with parameter identifiability using molecular data. Further work should permit incorporation of these issues into estimates of species trees provided they are rare across genes, but the assumptions required to model these factors may be very specific to the data set at hand. Thus, this first version of the Bayesian hierarchical model, based on coalescent theory, is a good starting point that can easily be generalized to use more realistic models of the coalescent process that are available.

We have discussed the effect of priors of the effective population size on species tree estimation. The estimate of the topology and divergence times of the species tree

is reasonably robust to changes in the prior of the effective population size; although naturally the estimate of the effective population size itself will be affected. The proposed model should be used to estimate ancestral population sizes only with extreme caution. The sensitivity of the estimates of the population sizes to the prior implies that either the prior is inappropriate or the information content in the data is low, or the likelihood is incorrect. Other empirical analyses suggest that ancestral population sizes are difficult to estimate under a wide variety of circumstances [116] [129] [58]. It is possible to use a non-informative prior or to combine the results from using different priors to make the result relatively robust to the prior. If this does not work, it may imply that there is not enough information in the data and accumulation of more genomic data may be needed. In this dissertation, we did not exploit the use of possible prior knowledge of the species trees. We used a birth-and-death process as the prior of species trees. Further work should incorporate other priors to see the effect on species tree and joint gene tree estimation. For example, the prior might be based on morphological or behavioral data in order to take such phylogenetically rich information.

An important byproduct of our model is the joint prior of gene trees. Our model formulates the correlation structure of gene trees across different genes through coalescent theory and the birth-death process. The correlation structure will be more realistic if our model includes key factors like horizontal transfer or gene duplication and uses a more appropriate prior for the species tree and population sizes. Thus, further research on model formulation is necessary. But the most important thing we stress here is the novel approach to estimating gene trees by employing the joint development of gene trees that are compatible with the species tree. Most current

approaches assume independent loci. It would be more reasonable to assume the loci are conditionally independent (given the species tree) but marginally dependent. Our method suggests that gene trees and species trees should be estimated simultaneously, and in that species tree estimation requires additional steps and considerations not traditionally included in phylogenetic analysis.

Using multiple allele datasets may improve the estimates of the species tree and ancestral population sizes. Unfortunately, the current version of the BEST procedure is unable to analyze the dataset with multiple alleles per species. One of the future directions of our research should thus be to extend our technique to handle the multiple allele data. This should be possible since formula for the distribution of gene trees given species trees in the coalescent model is still available for this more general case [102].

Our current technique for relaxing the molecular clock assumption is arbitrary. We might be able to overcome this limitation by treating mutation rates along lineages as parameters in the model and allowing different lineages to have different mutation rates. By doing this, we will use a non-clock-like gene tree to calculate the likelihood and use an ultrametric gene tree to calculate the prior. This technique could not only relax the clock assumption, but also will allow us to infer mutation rates along the different lineages.

BIBLIOGRAPHY

[1] H. Akaike. A new look at the statistical model identification. *IEEE Transactions on Automatic Control*, 19:716–723, 1974.

[2] B.L. Allen and M. Steel. Subtree transfer operations and their induced matrics on evolutionary trees. *Annuals of combinatorics*, 5:1–15, 2001.

[3] J.C. Avise, J.F. Shapiro, S.W. Daniel, C.F. Aquadroa, and R.A. Lansman. Mitochondrial dna differentiation during the speciation process in peromyscus. *Mol. Biol. Evol.*, 1:38–56, 1983.

[4] M. Barlow and B.G. Hall. Origin and evolution of the ampc β-lactamases of citrobacter freundii. *Antimicrob Agents Chemother*, 46:1190–1198, 2002.

[5] M. Barrett, M.J. Donoghue, and E. Sober. Against consensus. *Systematic Zoology*, 40(4):486–493, 1991.

[6] D.R. Brooks and D.A. McLennan. *Phylogeny, Ecology, and Behavior: A reasearch program in comparative biology*. University of Chicago press, 1991.

[7] J.J. Bull, J.P. Huelsenbeck, C.W. Cunningham, D.L. Swofford, and P.J. Waddell. Partitioning and combining data in phylogenetic analysis. *Systematic Biology*, 42:384–397, 1993.

[8] C. Cannings. The latent roots of certain markov chains arising in genetics: A new approach. i. haploid models. *Adv. Appl. Prob.*, 6:260–290, 1974.

[9] B.P. Carlin and S. Chib. Bayesian model choice via markov chain monte carlo. In *Research Report*, pages 93–006. University of Minnesota, 1993.

[10] B.P. Carlin and S. Chib. Bayesian model choice via markov chain monte carlo methods. *J. R.Statistics. Soc.B*, 57:473–484, 1995.

[11] B.P. Carlin and N.G. Polson. Inference for nonconjugate bayesian models using the gibbs sampler. *Canadian Journal of statistics*, 19:399–405, 1991.

[12] F.C. Chen and W.H. Li. Genomic divergences between humans and other hominoids and the effective population size of the common ancestor of humans and chimpanzees. *Am. J. Hum. Genet.*, 68:444–456, 2001.

[13] G. Coop and R.C. Griffiths. Ancestral inference on gene trees under selection. *Theoretical Population Biology.*, 66:219–232, 2004.

[14] J. Cracraft and D.P. Mindell. The early history of modern birds: A comparison of molecular and morphological evidence. In *Hierarchy of life*, pages 389–403. Elsevier, Amsterdam, 1989.

[15] A. de Queiroz. For consensus (sometimes). *Systematic Biology*, 42:368–372, 1993.

[16] A. de Queiroz, M.J. Donoghue, and J. Kim. Separate versus combined analysis of phylogenetic evidence. *Annual Review of Ecology and Systematics*, 26:657–681, 1995.

[17] J.H. Degnan and N.A. Rosenberg. Discordance of species trees with their most likely gene trees. *PLoS Genetics.*, 2:762–768, 2006.

[18] J.H. Degnan and L. Salter. Gene tree distributions under the coalescent process. *Evolution*, 59:24–37, 2005.

[19] P. Donnelly and S. Tavare. Coalescents and genealogical structure under neutrality. *Ann. Rev. Genet.*, 29:401–421, 1995.

[20] J.J. Doyle. Gene trees and species trees: Molecular systematics as one-character taxonomy. *Systematic Botany*, 17:144–163, 1992.

[21] D. Durand, B.V. Halldorsson, and B. Vernot. A hybrid micro-macroevolutionary approach to gene tree reconstruction. *Journal of Computational Biology*, 13:320–335, 2006.

[22] A.W.F. Edwards and L.L. Cavalli-Sforza. Reconstruction of evolution. *Annals of Human Genetics*, 27:105–106, 1963.

[23] S. Edwards and P. Beerli. Perspective: gene divergence, population divergence, and the variance in coalescence time in phylogeographic studies. *Evolution*, 54:1839–1854, 2000.

[24] N. Eldridge and J. Cracraft. *Phylogenetic Patterns and the Evolutionary Process.* Columbia University Press, 1980.

[25] W.J. Ewens. The sampling theory of selectively neutral alleles. *Theor. Pop. Biol.*, 3:87–112, 1972.

[26] W.J. Ewens. Population genetics theory - the past and the future. In *Mathematical and statistical developments of evolutionary theory*. Kluwer Academic Publishers, 1990.

[27] J.S. Farris. Methods for computing wagner trees. *Systematic Zoology*, 19:83–92, 1970.

[28] J.S. Farris. A probability model for inferring evolutionary trees. *Systematic Zoology*, 22:250–256, 1973.

[29] J.S. Farris, M. Kallersjo, A.G. Kluge, and C. Bult. Testing significance of incongruence. *Cladistics*, 10:315–319, 1994.

[30] J. Felsenstein. Maximum likelihood and minimum-steps methods for estimating evolutionary trees from data on discrete charaters. *Systematic Zoology*, 22:240–249, 1973.

[31] J. Felsenstein. Maximum likelihood estimation of evolutionary trees from continuous characters. *American Journal of Human Genetics*, 25:471–492, 1973.

[32] J. Felsenstein. Alternative methods of phylogenetic inference and their interrelationship. *Systematic Zoology*, 28:49–62, 1979.

[33] J. Felsenstein. Evolutionary trees from dna sequences: A maximum likelihood approach. *Journal of Molecular Evolution*, 17:368–376, 1981.

[34] J. Felsenstein. A likelihood approach to character weighting and what it tells us about parsimony and compatibility. *Biological Journal of the Linnean Society*, 16:183–196, 1981.

[35] J. Felsenstein. Parsimony in systematics: Biological and statistical issues. *Annual review of ecology and systematics*, 14:313–333, 1983.

[36] J. Felsenstein. Phylip (phylogeny inference package) version 3.6. distributed by the author. *Department of Genome Sciences, University of Washington, Seattle*, 2003.

[37] J. Felsenstein and G.A. Churchill. A hidden markov model approach to variation among sites in rate of evolution. *Molecular Biology and Evolution*, 13:93–104, 1996.

[38] R.A. Fisher. On the dominance ratio. *Proc. Roy. Soc. Edin.*, 42:321–431, 1922.

[39] W.M. Fitch. Toward defining the course of evolution: Minimum change for a specified tree topology. *Systematic Zoology*, 20:406–416, 1971.

[40] W.M. Fitch and E. Margoliash. Construction of phylogenetic trees. *Science*, 155:279–284, 1967.

[41] J. Fooden. Classification and distribution of living macaques. In *The macaques: studies in ecology, behavior, and evolution*, pages 1–9. Van Nostrand Reinhold, New York., 1980.

[42] Y.X. Fu and W.H. Li. Coalescing into the 21st century: An overview and prospects of coalescent theory. *Theoretical Population Biology*, 56:1–10, 1999.

[43] S.R. Gadagkar, M.S. Rosenberg, and S. Kumar. Inferring species phylogenies from multiple genes: Concatenated sequence tree versus consensus gene tree. *Journal of experimental zoology part B-molecular and developmental evolution*, 1:64–74, 2005.

[44] P.A. Goloboff. Analyzing large data sets in reasonable times: Solutions for composite optima. *Cladistics*, 15:415–428, 1999.

[45] M. Goodman, J Czelusniak, J.W. Moore, A.E. Romero-Herrera, and G. Matsuda. Fitting the gene lineage into its species lineage, a parsimony strategy illustrated by cladograms constructed from globin sequences. *Systematic Zoology*, 28:132–163, 1979.

[46] R.C. Griffiths. Lines of descent in the diffusion approximation of neutral wright-fisher models. *Theor. Pop. Biol.*, 17:37–50, 1980.

[47] R.C. Griffiths and S. Tavare. Computational methods for the coalescent. In *Population Genetics and Human Evolution.*, pages 165–182. Springer Verlag, New York., 1997.

[48] P.H. Harvey and M.D. Pagel. *Comparative method in evolutionary biology.* Oxford University Press, 1991.

[49] M. Hasegawa, H. Kishino, and T. Yano. Dating of the human-ape splitting by a molecular clock of mitochondrial dna. *Journal of Molecular Evolution*, 26:132–147, 1985.

[50] W.K. Hastings. Monte carlo sampling methods using markov chains and their applications. *Biometrika*, 57:97–109, 1970.

[51] R. R. Hudson and N. L. Kaplan. Statistical properties of the number of recombination events in the history of a sample of dna sequences. *Genetics*, 111:147–164, 1985.

[52] R.R. Hudson. Estimating genetic variability with restriction endonucleases. *Genetics*, 100:711–719, 1982.

[53] R.R. Hudson. Gene genealogies and the coalescent process. In *Oxford Surveys in Evolutionary Biology*, pages 1–44. Oxford University Press, 1991.

[54] R.R. Hudson. The how and why of generating gene genealogies. In *Mechanisms of Molecular Evolution*, pages 23–26. Sinauer., 1992.

[55] J.P. Huelsenbeck, J.J. Bull, and C.W. Cunningham. Combining data in phylogenetic analysis. *Trends Ecol Evol*, 11:152–158, 1995.

[56] J.P. Huelsenbeck and D.M. Hills. Success of phylogenetic methods in the four-taxon case. *Systematic Biology*, 42:247–264, 1993.

[57] J.P. Huelsenbeck, D.L. Swofford, C.W. Cunningham, J.J. Bull, and P.J. Waddell. Is character weighting a panacea for the problem of data heterogeneity in phylogenetic analysis? *Systematic Biology*, 43(9):288–291, 1994.

[58] W.B. Jennings and S.V. Edwards. Speciational history of australian grass finches (poephila) inferred from thirty gene trees. *Evolution*, 59(9):2033–2047, 2005.

[59] T.H. Jukes and C.R. Cantor. Evolution of protein molecules. In *Mammalian Protein Metabolism*, pages 21–132. Academic Press, 1969.

[60] N.L. Kaplan, T. Darden, and R.R. Hudson. The coalescent process in models with selection. *Genetics*, 120:819–829, 1988.

[61] S. Karlin and J. McGregor. Addendum to a paper of w. ewens. *Theoret. Popn. Biol.*, 3:113–116, 1972.

[62] M. Kimura. A simple model for estimating evolutionary rates of base substitutions through comparative studies of nucleotide sequences. *Journal of Molecular Evolution*, 16:111–120, 1980.

[63] J.F.C. Kingman. Mathematics of genetic diversity. In *CBMSNSF Regional Conference Series in Applied Mathematics*, page 34. Society for Industrial and Applied Mathematics, Philadelphia, Pennsylvania., 1980.

[64] J.F.C. Kingman. The coalescent. *Stoch. Proc. Applns.*, 13:235–248, 1982.

[65] J.F.C. Kingman. Exchangeability and the evolution of large populations. In *Exchangeability in Probability and Statistics*, pages 97–112. North-Holland Publishing Company, 1982.

[66] J.F.C. Kingman. On the genealogy of large populations. *J. Appl. Prob.*, 19A:27–43, 1982.

[67] H. Kishino and M. Hasegawa. Evaluation of the maximum likelihood estimate of the evolutionary tree topologies from dna sequence data, and the branching order in hominoidea. *Journal of Molecular Evolution*, 29:170–179, 1989.

[68] A.G. Kluge. A concern for evidence and a phylogenetic hypothesis of relationships among epicrates(boidae,serpentes). *Systematic Zoology*, 38:7–25, 1989.

[69] A.G. Kluge and J.S Farris. Quantitative phyletics and the evolution of anurans. *Systematic Zoology*, 18:1–32, 1969.

[70] A.G. Kluge and A.J Wolf. Cladistics: what's in a word. *Cladistics*, 9:183–199, 1993.

[71] L.S. Kubatko and J.H. Degnan. Inconsistency of phylogenetic estimates from concatenated data under coalescence. *Systematic Biology In revision*, 2006.

[72] W.J. Le Quesne. The uniquely evolved character concept and its cladistic application. *Systematic Zoology*, 23:513–517, 1974.

[73] S. Li, D.K. Pearl, and H. Doss. Phylogenetic tree construction using markov chain monte carlo. *Journal of the American Statistical Association*, 95:493–508, 2000.

[74] W.H. Li and Y.X. Fu. Coalescent theory and its applications in population genetics. In *Statistics in Genetics*, pages 45–79. Springer Verlag, 1999.

[75] M. Ludy. Applications of the annealing algorithm to combinatorial problems in statistics. *Biometrika*, 72:91–98, 1985.

[76] J.F. Maa, D.K. Pearl, and R. Bartoszytiski. Reducing multidimensional two-sample data to one-dimensional interpoint distances. *The annals of statistics*, 24:1069–1074, 1996.

[77] D.R. Maddison. *Phylogenetic inference of historical pathways and models of evolutionary change*. Harvard University, 1990.

[78] W.P. Maddison. Molecular approaches and the growth of phylogenetic biology. In *Molecular zoology: Advances, strategies, and protocols*, pages 47–63. Wiley-Liss, 1996.

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

[80] W.P. Maddison, M.J. Donoghue, and D.R. Maddison. Outgroup analysis and parsimony. *Systematic Zoology*, 33:83–103, 1984.

[81] W.P. Maddison and L.L. Knowles. Inferring phylogeny despite incomplete lineage sorting. *Systematic Biology*, 55:21–30, 2006.

[82] B. Mau and M.A. Newton. Phylogenetic inference for binary data on dendrograms using markov chain monte carlo. *Journal of Computational and Graphical Statistics*, 6:122–131, 1997.

[83] X.L. Meng and W.H. Wong. Simulating ratios of normalizing constants via a simple identity. In *Technique report 365*. University of Chicago, 1993.

[84] N. Metropolis, A.W. Rosenbluth, M.N. Rosenbluth, A.H. Teller, and E. Teller. Equation of state calculations by fast computing machines. *Journal of Chemical Physics*, 21:1087–1092, 1953.

[85] V. Minin, Z. Abdo, P. Joyce, and J. Sullivan. Performance-based selection of likelihood models for phylogeny estimation. *Systematic Biology*, 52:674–683, 2003.

[86] M.M. Miyamoto and W.M. Fitch. Testing species phylogenies and phylogenetic methods with congruence. *Systematic Biology*, 44:64–76, 1995.

[87] W.S. Moore. Inferring phylogenies from mtdna variation: mitochondrial gene trees vs. nuclear gene trees. *Evolution*, 49:718–726, 1995.

[88] S. Nee, R.M. May, and P.H. Harvey. The reconstructed evolutionary process. *Philosophical Transactions of the Royal Society of London, series B*, 344:77–82, 1995.

[89] C. Neuhauser and S.M. Krone. The genealogy of samples in models with selection. *Genetics*, 145:519–534, 1997.

[90] C. Neuhauser and S. Tavare. The coalescent. In *Encyclopedia of Genetics.*, pages 392–397. Academic Press, 2001.

[91] M.A. Newton and A.E. Raftery. Approximate bayesian inference by the weighted likelihood bootstrap. *J. R.Statistics. Soc.B*, 56:3–48, 1994.

[92] R. Nielsen, J.L. Mountain, J.P. Huelsenbeck, and M. Slatkin. Maximum likelihood estimation of population divergence times and population phylogenies under the infinite sites model. *Theoretical Population Biology*, 53:143–151, 1998.

[93] R. Nielsen, J.L. Mountain, J.P. Huelsenbeck, and M. Slatkin. Maximum-likelihood estimation of population divergence times and population phylogeny in models without mutation. *Evolution*, 52:669–677, 1998.

[94] K.C. Nixon. The parsimony ratchet, a new method for rapid parsimony analysis. *Cladistics*, 15:407–414, 1999.

[95] K.C. Nixon and J.M. Carpenter. On simultaneous analysis. *Cladistics*, 12:221–241, 1996.

[96] M. Nordborg. Coalescent theory. In *Handbook of Statistical Genetics*, pages 179–208. John Wiley and Sons, Inc, 2001.

[97] J.A.A. Nylander, F. Ronquist, J.P. Huelsenbeck, and J.L. Nieves Aldrey. Bayesian phylogenetic analysis of combined data. *Systematic Biology*, 53:47–67, 2004.

[98] R.D.M. Page. Genetree: Comparing gene and species phylogenies using reconciled trees. *Bioinformatics*, 14:819–820, 1998.

[99] P. Pamilo and M Nei. Relationships between gene trees and species trees. *Molecular Biology and Evolution*, 5:568–583, 1988.

[100] D. Penny and M.D. Hendy. Turbotree: A fast algorithm for minimal trees. *Computer applications in the biosciences*, 3:183–187, 1987.

[101] D. Posada and K.A. Crandall. Modeltest: testing the model of dna substitution. *Bioinformatics*, 14(9):817–818, 1998.

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

[103] D.F. Robinson and L.R. Foulds. Comparison of phylogenetic trees. *Math. Biosci.*, 53:131–147, 1981.

[104] A.G. Rodrigo, M. Kellyborges, P.R. Bergquist, and P.L. Bergquist. A randomization test of the null hypothesis that two cladograms are sample estimates of a parametric phylogenetic tree. *New Zealand J. Bot.*, 31(9):257–268, 1993.

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

[106] F. Ronquist, Huelsenbeck J.P., and L. van der Mark. Mrbayes3 manual. *online*, 2005.

[107] S. Rosenkranz. The bayes factor for model evaluation in a hierarchical poisson model for area counts. In *Ph.D dissertation*, pages 93–006. University of Washington, 1992.

[108] M. Ruvolo. Molecular phylogeny of the hominoids: Inferences from multiple independent dna sequence data sets. *Molecular Biology and Evolution*, 14:248–265, 1997.

[109] L. Salter and D.K. Pearl. Stochastic search strategy for estimation of maximum likelihood phylogenetic trees. *Systematic Biology*, 50:7–17, 2001.

[110] M.P. Simmons, C.D. Bailey, and K.C. Nixon. Phylogeny reconstruction using duplicate genes. *Molecular Biology and Evolution*, 17(4):469–473, 2000.

[111] M Slatkin and J.L. Pollack. The concordance of gene trees and species trees at two linked loci. *Genetics.*, 172:1979–1984, 2006.

[112] J.B. Slowinski and R.D.M. Page. How should species phylogenies be inferred from sequence data? *Systematic Biology*, 48(4):814–825, 1999.

[113] D.L. Swofford. When are phylogeny estimates from molecular and morphological data incongruent? In *Phylogenetic Analysis of DNA Sequences*, pages 295–333. Oxford University Press, 1991.

[114] D.L Swofford and G.J Olsen. Phylogeny reconstruction. In *Molecular Systematics*, pages 411–501. Sinauer Associates, 1990.

[115] F. Tajima. Evolutionary relationship of dna sequences in finite populations. *Genetics*, 105:437–460, 1983.

[116] N. Takahata. Gene genealogy in three related populations: Consistency probability between gene and population trees. *Genetics*, 122:957–966, 1989.

[117] S. Tavare. Line-of-descent and genealogical processes, and their applications in population genetics models. *Theoret. Popn. Biol.*, 26:119–164, 1984.

[118] S. Tavare. Some probabilistic and statistical problems in the analysis of dna sequences. In *Lectures on Mathematics in the Life Sciences*, pages 57–86. American Mathematical Society, 1986.

[119] S. Tavare. Calibrating the clock: using stochastic processes to measure the rate of evolution. In *Calculating the secrets of life*, pages 114–152. National Academy Press, 1993.

[120] S. Tavare. *Ancestral Inference in Population Genetics*. Springer Lecture Notes in Mathematics, 2003.

[121] A.R. Templeton. Phylogenetic inference from restriction endonuclease cleavage site maps with particular reference to the evolution of apes and humans. *Evolution*, 37:221–244, 1983.

[122] A.J. Tosi, J.C. Morales, and D.J. Melnick. Paternal, maternal, and biparental molecular markers provide unique windows onto the evolutionary history of macaques monkeys. *Evolution*, 57(6):1419–1435, 2003.

[123] A. Wald. Note on the consistency of the maximum likelihood estimate. *Ann. Math. Statist.*, 29:595–601, 1949.

[124] G.A. Watterson. On the number of segregating sites in genetical models without recombination. *Theoret. Popn. Biol.*, 7:256–276, 1975.

[125] S. Wright. Evolution in mendelian populations. *Genetics*, 16:97–159, 1931.

[126] S. Wright. Size of population and breeding structure in relation to evolution. *Science*, 87:430–431, 1938.

[127] C.I. Wu. Inferences of species phylogeny in relation to segregation of ancient polymorphisms. *Genetics*, 127:429–435, 1991.

[128] Z. Yang. On the estimation of ancestral population sizes. *Genet. Res.*, 69:111–116, 1997.

[129] Z. Yang and B. Rannala. Bayesian phylogenetic inference using dna sequences: A markov chain monte carlo method. *Molecular Biology and Evolution*, 14:717–724, 1997.

[130] Z. Yang and B. Rannala. Likelihood and bayes estimation of ancestral population sizes in hominoids using data from multiple loci. *Genetics*, 162:1811–1823, 2002.