To be submitted to *Systematic Biology* *July 18, 2006*

SPECIES TREES FROM GENE TREES

Species Trees from Gene Trees: reconstructing Bayesian posterior distributions of a species phylogeny using estimated gene tree distributions

Liang Liu[1,2] and Dennis K. Pearl[1]

[1]*Department of Statistics, The Ohio State University,*
Columbus, OH 43210-1293

[2]Corresponding author and address for correspondence:

Liang Liu
Department of Statistics,
The Ohio State University
1958 Neil Avenue
Cockins Hall
Columbus, OH 43210-1247
liu.473@osu.edu

Ph: 614-292-1093
FAX: 614-292-2096

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 paper, we 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. Our 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 estimate gene trees, species trees, ancestral population sizes, and species divergence times simultaneously. Our 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. We apply the method to two 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.

Keywords: species tree, gene tree, coalescent theory, MCMC, Bayesian hierarchical model.

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 (Maddison, 1997; Slowinski, Maddison and Knowles, 2006). Thus, in fact, molecular 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 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 (1989, 1993) claimed 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 (Kluge, 1989; Kluge and Wolf, 1993; Nixon and Carpenter, 1996) 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 (Slowinski and Page, 1999) since it assumes that nucleotides are independent estimates of the species tree so 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 (Kubatko and Degnan, 2006; Degnan and Rosenberg, 2006). 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 (Page, 1998; Pamilo, 1988). This idea is based on Doyle's (1992) concept that nucleotides are characters of gene trees, whereas gene trees are characters of species trees (Maddison, 1997). 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 (Durand et al, 2006; Coop and Griffiths, 2004). 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 (Page, 1998; Simmons et al., 2000) and consensus tree methods (Bull et al., 1993; de Queiroz, 1993; Rodrigo et al., 1993; Huelsenbeck et al., 1994) both ignore the errors in gene tree estimation and generally assume that gene trees are estimated with perfect certainty. In these approaches, 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 (2005) have derived the probability distribution for the topology of gene trees given the species tree. Slatkin and Pollack (2006) 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 (Nielsen et al, 1998a; Nielsen et al, 1998b).

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.

This 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.

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, $\theta = 4N_e\mu$.

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

$$f(S,\theta \mid D) = \frac{1}{f(D)} \int_{\Lambda} \int_{\mathbf{G}} f(D \mid \mathbf{G},\Lambda) * f(\Lambda) * f(\mathbf{G} \mid S,\theta) * f(\theta) * f(S) d\mathbf{G}\, d\Lambda .$$

Our Bayesian hierarchical structure consists of modeling the following components: f(D|**G**, Λ), f(Λ), f(**G**|S, θ), f(S), and f(θ), each of which is explained below.

(1) f(D|**G**, Λ).

Markovian models dominate the likelihood based literature for both nucleotide and amino acid substitution (Felsenstein, 2004b). 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**, Λ) can be appropriately described. The quantity f(D|**G**, Λ) 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 (Posada and Crandall, 1998) or information theory (Minin et al., 2003).

(2) f(Λ).

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 (Ronquist, Huelsenbeck, and van der Mark, 2005).

(3) $f(\mathbf{G} \mid S,\theta)$.

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 $\theta = 4\mu N_e$ 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 $t_{n+1},...,t_m$ 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:

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

Thus $f(G \mid S,\theta)$ 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 $f(\mathbf{G} \mid S,\theta)$. 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. So the cumulative node-to-tip branch lengths in the gene trees are always longer 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 below, we assume their effective population sizes are one-fourth that of autosomal markers.

(4) f(θ).

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

(5) f(S)

We use a birth-and-death process (Nee, May, and Harvey, 1994) 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 (n), the joint density of the topology (T) and branch lengths (τ) of a particular species tree is (Yang and Rannala, 1997):

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

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

Mutation rate variation among loci may influence the estimation of ancestral population sizes (Yang 1997; Chen and Li 2001). Nevertheless, if the ratios of rates between loci are known, we can incorporate them in the likelihood calculation (Yang, 1997). 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.

COMPUTATIONAL ALGORITHM

The entire species tree estimation procedure consists of three steps.

Step 1 (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.

Step 2: 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.

Step 3: 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,

$$f(S,\theta \mid D) = \int_{\Lambda}\int_{G} f(S,\theta,G,\Lambda \mid D)dGd\Lambda$$

$$= \frac{1}{f(D)}\int_{\Lambda}\int_{G} f(D \mid G,\Lambda) * f(\Lambda) * f(G \mid S,\theta) * f(\theta) * f(S)dGd\Lambda$$

$$= \frac{1}{f(D)}\int_{\Lambda}\int_{G} f(D \mid G,\Lambda) * f(\Lambda) * \frac{f(G)}{f(G)} * f(G \mid S,\theta) * f(\theta) * f(S)dGd\Lambda$$

$$= \frac{1}{f(D)}\int_{\Lambda}\int_{G} f(D \mid G,\Lambda) * f(\Lambda) * f(G)\frac{f(G \mid S,\theta) f(\theta) f(S)}{f(G)}dGd\Lambda$$

$$= \int_{G} f(G \mid D) f(S,\theta \mid G)dG . \qquad (1)$$

The posterior of species tree and population sizes given data, $f(S,\theta \mid D)$, is the posterior of species tree and θ given gene trees $f(S,\theta \mid \mathbf{G})$ weighted by $f(\mathbf{G} \mid D)$. 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 integration of f(**G**|S) with respect to the species tree (topology and branch lengths) and population size θ, namely,

$$f(\mathbf{G}) = \int_{\theta} \int_{S} f(\mathbf{G} \mid S, \theta) f(S) f(\theta) dS d\theta .$$

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 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 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 decides 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 $\mathbf{G}_1$, $\mathbf{G}_2$, ..., $\mathbf{G}_N$ along with the associated approximate priors K($\mathbf{G}_1$), K($\mathbf{G}_2$), ..., K($\mathbf{G}_N$) 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 was used to define the prior distribution of species trees (Nee, May, and Harvey, 1994) and the likelihood is defined by coalescent theory via

Rannala and Yang's formula. The movement strategy employed 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 $f(S\,|\,\mathbf{G}_i)$ for each of the gene tree vectors $\mathbf{G}_1$, $\mathbf{G}_2$, ..., $\mathbf{G}_N$ 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:

$$\begin{aligned} f(S,\theta\,|\,D) &= \int_{\mathbf{G}} f(\mathbf{G}\,|\,D) f(S,\theta\,|\,\mathbf{G}) d\mathbf{G} \\ &= \int_{\mathbf{G}} \left[K(\mathbf{G}\,|\,D)\frac{f(\mathbf{G})}{K(\mathbf{G})}\right] f(S,\theta\,|\,G)dG. \end{aligned}$$

The ith sample from step 1 gives the value for K($\mathbf{G}_i$|D). We need to multiply this by $\dfrac{f(\mathbf{G}_i)}{K(\mathbf{G}_i)}$ in order to align it with $f(\mathbf{G}_i\,|\,D)$ and produce samples from the true posterior.

Despite the fact that $f(\mathbf{G}_i)$ is not known, we can apply the harmonic mean technique (Newton and Raftery, 1994) to estimate it. In particular, we have

$$\frac{1}{f(\mathbf{G}_i)} \propto \int_{speciestree} \frac{1}{f(\mathbf{G}_i)} f(S)dS \quad = \int_{speciestree} \frac{1}{f(\mathbf{G}_i\,|\,S)} \frac{f(\mathbf{G}_i\,|\,S)}{f(\mathbf{G}_i)} f(S)dS$$

$$= \int_{speciestree} \frac{1}{f(\mathbf{G}_i\,|\,S)} f(S\,|\,\mathbf{G}_i)dS\,,$$

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 $\mathbf{G}_i$. Thus, a consistent estimate of $f(\mathbf{G}_i)$ is given by $\hat{f}(\mathbf{G}_i) = \hat{\alpha}_i \left(\sum_{j=1}^{k} \frac{1}{f(\mathbf{G}_i\,|\,S_j)} \right)^{-1}$ using the samples S$_1$, ..., S$_k$ from $f(S\,|\,\mathbf{G}_i)$ 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

$$\left[(S_1,\mathbf{G}_1),(S_2,\mathbf{G}_1),...,(S_k,\mathbf{G}_1)\right],\left[(S_1,\mathbf{G}_2),(S_2,\mathbf{G}_2),...,(S_k,\mathbf{G}_2)\right],...,\left[(S_1,\mathbf{G}_N),(S_2,\mathbf{G}_N),...,(S_k,\mathbf{G}_N)\right]$$

where the block of pairs $\left[(S_1,\mathbf{G}_i),(S_2,\mathbf{G}_i),...,(S_k,\mathbf{G}_i)\right]$ is given total weight

$$\frac{\hat{f}(\mathbf{G}_i)}{K(\mathbf{G}_i)}\left(\sum_{i=1}^{N}\frac{\hat{f}(\mathbf{G}_i)}{K(\mathbf{G}_i)}\right)^{-1}.$$

THEORETICAL RESULTS

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 (Nylander, et al., 2004). 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 (Barlow and Hall, 2002). 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 the DNA sequences given gene trees is just the product of the likelihood for each gene:

$$L^{BCT} = f(D\,|\,\mathbf{G}) = f(D_1...D_K\,|\,G_1...G_K) = \prod_{i=1}^{K} f_i(D_i\,|\,G_i).$$

The prior of the gene trees for different genes is then the product of the priors for each gene:

$$\text{Prior}^{\text{BCT}} = f(\mathbf{G}) = f(G_1 ... G_K) = \prod_{i=1}^{K} f_i(G_i).$$

For the Bayesian concatenation method, the likelihood is based on the additional assumption that all the genes arise from the same tree G*:

$$\text{L}^{\text{BCM}} = f(D_1 ... D_k \mid G^*) = \prod_{i=1}^{k} f_i(D_i \mid G^*).$$

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

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

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 is assumed that all the gene trees are the same). 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 possibly 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 (Kubatko and Degnan, 2006; Degnan and Rosenberg, 2006). Of course, neither method facilitates estimation of important parameters in the evolutionary history of species such as population sizes or speciation times. Speciation times here are distinct from gene divergence times due to the coalescent process (Edwards and Beerli, 2000).

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 a joint prior instead of an independent prior for gene trees. However we don’t assume that 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.

DATA ANALYSIS

Australian Finch Data

We first apply the new method to a multilocus nucleotide dataset from birds recently published by Jennings and Edwards (2005). They obtained the allelic data of 30 loci (Pa-01…Pa-30) from one individual per population of *Poephila acuticauda, P.hecki, P.cincta.* They also included sequences from a more distant relative, the zebra finch (*P.guttata*), as outgroup. A total of 30 anonymous loci were developed ranging in size from 216 to 825 bp. They performed a "four-gamete test" and the result showed that the overall incidence of intralocus recombination in the data appears is very 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.

(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 (Hasegawa, Kishino and Yano, 1985) was selected as the substitution model that best fit the data according to an hierarchical likelihood ratio test. Since the position of species 4 is fixed as the outgroup, there are only three possible topologies, (2,(1,3)), (3,(1,2)), and (1,(2,3)). From Table 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 the 30 genes is 0.434. The corresponding probabilities for the other two possible trees are 0.205 and 0.361. Thus the tree (3,(1,2)) has slightly 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) Bayesian estimation of the species tree for the concatenation method.

In this case the multilocus sequences are concatenated into a single sequence. The concatenated data was analyzed in MrBayes with an HKY85 substitution model. The prior for the topology was taken to be a uniform distribution, and branch lengths were assumed to be independently distributed exponentials. The resulting estimated topology of the species tree is in Table 1. It matches the majority-rule tree in (1) and its posterior probability is essentially one.

(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 an 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 reconstruction program (BEST). Three different priors of effective population sizes were used to evaluate the effect of the prior on the posterior distribution. The priors are exponential distributions with means 1, 0.1, 0.0072, and 0.00072. The median 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 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)

The same species tree with strong support is estimated regardless which of the priors we use (Table 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 1).

To compare the BEST method with the concatenation method, we estimate the Bayes factor using the harmonic mean of the likelihood. Although the harmonic mean method can be somewhat unstable and sensitive to the lowest value of the likelihood, it works here since the likelihoods for the two different methods are well separated (Figure 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 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 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 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

$$f(\mathbf{G}) = f(G_1 ... G_K) = \int_S \int_\theta f(G_1 | S,\theta) * ... * f(G_K | S,\theta) * f(\theta) * f(S) dS d\theta$$

where $f(G_i | S,\theta)$ follows coalescent theory. $f(G_1 ... G_k)$ 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 1. There are 22 genes supporting the tree (3,(,1,2)) in Table 1, compared with only 15 genes when the independent prior was used (Table 1). 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 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 in $f(\mathbf{G})$ 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 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 changed even if the joint prior is used. A similar pattern is seen with Pa-21, 16 and 30 in Table 1.

Our estimate of the species tree agrees with the assumed species phylogeny found by Jennings and Edwards (2005). 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.

Analysis of multilocus Macaque data

Tosi and Morales (2003) 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 19 species, including one outgroup taxa (*T. Gelada*), for which there is data on all four "genes." We randomly chose an allele from each species. 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.

(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 (Tosi and Morales, 2003). 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 length 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 (Miyamoto and Fitch 1995; Moore 1995; Ruvolo 1997).

Fooden defined four species groups for macaques according to distinct forms of male reproductive anatomy (Fooden, 1980). 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 relative high posterior probability (Figure 2). The species in the other three groups are not well resolved. This indicates 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 D(T1,T2)be the symmetric distance between two random trees T1 and T2 (Robinson and Foulds, 1981). Table 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. This 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 if the posterior distributions of gene trees using two different priors are different. To test if two distributions are different, we introduce a theorem by Maa et al. (1996).

We want to test the hypothesis H: $F_1=F_2$, where F_1 and F_2 are two distributions such as the two posterior tree distributions examined here. Let X_1 and X_2 be independent and identically distributed random draws from F_1 and independent of Y_1 and Y_2 from F_2. Take D(.,.) to be any appropriately chosen distance function. The theorem posits that $F_1=F_2$ if and only if $D(X_1 , X_2) = D(Y_1 , Y_2)= D(X_1 , Y_1)$ in distribution.

To apply the theorem, we calculated the three distances (two within group distances and the between groups distance) for each gene in Table 4. 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 data set.

(3) Sensitivity analysis.

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 S_1, S_2, S_3, and S_4 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 S_i 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 a single outlier gene.

DISCUSSION

Different genes may have different mechanisms of evolution and support different phylogenetic relationships. Mixing this diverse set of circumstances together through concatenation is then inappropriate (Mossel, 2005) and may lead to difficulties in the performance of algorithms that do not recognize the problem explicitly. This work defines the correlation between gene trees through their common species tree but then, conditional on this species tree, allows for a completely independent evolutionary processes for each gene.

The Bayesian hierarchical model we have employed adopts coalescent theory to formulate the distribution of gene trees given the species trees. A simulation study has found that increasing the number of loci gives more accurate estimate of the species tree under our assumption that deep coalescence is the only reason for the conflicts between the gene tree and species tree (Maddison and Knowles, 2006). 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 diagnostics of model adequacy and into estimates of species trees provided they are rare across genes. 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 robust 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 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 (Takahata 1989; Yang 1997; Jennings and Edwards 2005).

In this paper, 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 the estimation of species tree and joint gene tree distributions. For example, we might use a more informed prior centered on the distribution arising from other types of data such as behavioral or morphologic information or we might use a less informed prior in which all species trees are equally likely within some bounds. Of course this choice should be left to the individual investigator but it is important to understand the relative contribution it plays in determining the posterior compared with the information in the data. If it is desired to have the phylogeny completely resolved by the data alone, then the accumulation of more genomic data may be needed.

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 that likelihood-based species tree estimation requires an explicit model and corresponding algorithms not traditionally included in phylogenetic analysis.

ACKNOWLEGEMENTS

We would like to thank Scott Edwards, Bryan Jennings, and Anthony Tosi for providing the sequences data. Special thanks to Scott Edwards for his insightful comments and suggestions for improving the model. This work was partially supported by grant NSF DMS-0112050 to the Mathematical Biosciences Institute.

References

Altekar, G., S. Dwarkadas, J. P. Huelsenbeck., and F. Ronquist. 2004. Parallel metropolis-coupled markov chain monte carlo for bayesian phylogenetic inference. *Bioinformatics.* 20:407-415.

Barlow, M., and B.G. Hall. 2002. Origin and evolution of the AmpC β-Lactamases of citrobacter freundii. *Antimicrob Agents Chemother*. 46: 1190–1198.

Bull J. J., Huelsenbeck, J. P., Cunningham, C. W., Swofford, D. L., and P. J. Waddell. 1993. Partitioning and combining data in phylogenetic analysis. *Syst. Biol.* 42:384–397.

Chen, F. C., and W. H. Li. 2001. 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.

Coop, G., and R. C. Griffiths. 2004. Ancestral inference on gene trees under selection. *Theor. Popul. Biol.* 66: 219-232.

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

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

de Queiroz A. 1993. For consensus (sometimes). *Syst Biol.* 42:368–372.

Doyle, J. J. 1992. Gene trees and species trees: molecular systematics as one-character taxonomy *Syst. Bot.* 17:144-163.

Durand D., Halldorsson, B. V., and B. Vernot. 2006. A hybrid micro-macroevolutionary approach to gene tree reconstruction. *J. Comput. Biol.* 13: 320-335.

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

Felsenstein, J. 2004a. *Inferring Phylogenies*. Sinauer Associates, Inc. 664 pages.

Felsenstein, J. 2004b. PHYLIP (Phylogeny Inference Package) version 3.6. *Distributed by the author. Department of Genome Sciences, University of Washington, Seattle.*

Fooden, J. 1980. Classification and distribution of living macaques. D.G Lindburg, ed. The macaques: studies in ecology, behavior, and evolution. Van Nostrand Reinhold, New York.

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

Huelsenbeck, J. P., Swofford, D. L., Cunningham, C. W., Bull, J. J., and P. J. Waddell. 1994. Is character weighting a panacea for the problem of data heterogeneity in phylogenetic analysis? *Syst Biol.* 43:288–291.

Huelsenbeck, J. P, and F. Ronquist. 2001. Mrbayes: Bayesian inference of phylogeny. *Bioinformatics.* 17:754-755.

Jennings, W. B., and S.V. Edwards. 2005. Speciational history of Australian grass finches (*Poephila*) inferred from thirty gene trees. *Evolution*. 59: 2033-2047.

Kluge, A. G.. 1989. A concern for evidence and a phylogenetic hypothesis of relationships among Epicrates (Boidae,Serpentes). *Syst.Zool.* 38:7-25.

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

Kubatko, L. S., and J. H. Degnan. 2006. Inconsistency of phylogenetic estimates from concatenated data under coalescence. *Syst. Biol.* In revision.

Maa, J. F., Pearl, D. K., and R. Bartoszytiski. 1996. Reducing multidimensional two-sample data to one-dimensional interpoint distances. *Ann. Stat.* 24:1069-1074.

Maddison, W. P. 1997. Gene trees in species trees. *Syst. Biol.* 46: 523-536.

Maddison, W. P., and L. L. Knowles. 2006. Inferring phylogeny despite incomplete lineage sorting. *Syst Biol.* 55:21-30.

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

Miyamoto, M. M., and W. M. Fitch. 1995. Testing species phylogenies and phylogenetic methods with congruence. *Syst. Biol.* 44:64-76.

Moore, W. S. 1995 Inferring phylogenies from mtDNA variation: mitochondrial gene trees vs. nuclear gene trees. *Evolution.* 49:718-26.

Mossel, E., and E. Vigoda. 2005. Phylogenetic MCMC algorithms are misleading on mixtures of trees. Science. 309:2207-2209.

Nee, S., May, R. M. and P. H. Harvey. 1994. The reconstructed evolutionary process. *R. Soc. Lond. Proc. Ser. B* 344:77-82

Newton, M. A. and A. E. Raftery. 1994. Approximate Bayesian inference by the weighted likelihood bootstrap (with discussion). *J. Roy. Statistical Society, Series B* 56: 3–48.

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

Nielsen, R. 1998. Maximum likelihood estimation of population divergence times and population phylogenies under the infinite sites model. *Theor. Popul. Biol.* 53:143-151.

Nixon, K. C. and J. M. Carpenter. 1996. On simultaneous analysis. *Cladistics*. 12:221-241.

Nylander, J. A. A., Ronquist, F., Huelsenbeck, J. P., and J. L. Nieves Aldrey. 2004. Bayesian phylogenetic analysis of combined data. *Syst Biol.* 53:47-67.

Page, R. D. M. 1998. GeneTree: Comparing gene and species phylogenies using reconciled trees. *Bioinformatics.* 14:819-820.

Pamilo, P., and M. Nei. 1988. Relationships between gene trees and species trees. *Mol. Biol. Evol.* 5:568-583.

Posada, D. and K. A. Crandall. 1998. MODELTEST**:** testing the model of DNA substitution. *Bioinformatics* 14: 817-818.

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

Robinson, D. F. and L. R. Foulds. 1981. Comparison of phylogenetic trees. *Math. Biosci.* 53, 131-147.

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

Ronquist, F., and J. P. Huelsenbeck. 2003. MRBAYES 3: Bayesian phylogenetic inference under mixed models. *Bioinformatics.* 19:1572-1574.

Ronquist, F., Huelsenbeck, J. P. and P. van der Mark. 2005. *MrBayes3 manual*. On-line at http://mrbayes.csit.fsu.edu/mb3.1_manual.pdf

Ruvolo, M. 1997. Molecular phylogeny of the hominoids: Inferences from multiple independent DNA sequence data sets. *Mol. Biol. Evol*. 14:248-265.

Simmons, M. P., Bailey, C. D., and K. C. Nixon. 2000. Phylogeny reconstruction using duplicate genes. *Mol. Biol. Evol.* 17:469-473.

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

Slowinski, J. B. and R. D. M. Page. 1999. How should species phylogenies be inferred from sequence data? *Syst. Biol.* 48:814-825.

Takahata, N. 1989. Gene genealogy in three related populations: consistency probability between gene and population trees. *Genetics.* 122:957-966.

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

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

Yang, Z. and B. Rannala. 1997. Bayesian phylogenetic inference using DNA sequences: a markov chain monte carlo method. *Mol. Biol. Evol.* 14:717-724.

Table1. The posterior distributions of gene trees and species trees in the finch data set. There are only three possible topologies for 3 species. The posterior probability for each topology is listed for each gene. The highest value is highlighted. The row of labeled concatenation is the posterior probabilities of species trees for concatenation method. The Bayesian estimate is the second topology (3,(1,2)) with probability 1. The rows of joint priors are the posterior distributions of species trees for the BEST method with different priors for θ, gamma(1,139), gamma(1,1389), gamma(1,10) and gamma(1,1).

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

Table2. 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 in the column2 and column3. (1,2) represents the ancestral population of species1 and species2. (1,2,3) is the ancestral population of species1, species2 and species3.

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)

Table3. 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 specie 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, 2004) using the symmetric distance measure (Robinson.and Foulds, 1981).

	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

Table4. 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.

	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

Table5. 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.

S1 (without Y-Chromosome)	S2 (without mtDNA)	S3 (without C4 Intron 9)	S4 (without IRBP Intron 3)
0.683 (0.597,0.768)	0.682 (0.607, 0.757)	0.664 (0.583, 0.746)	0.663 (0.580, 0.747)

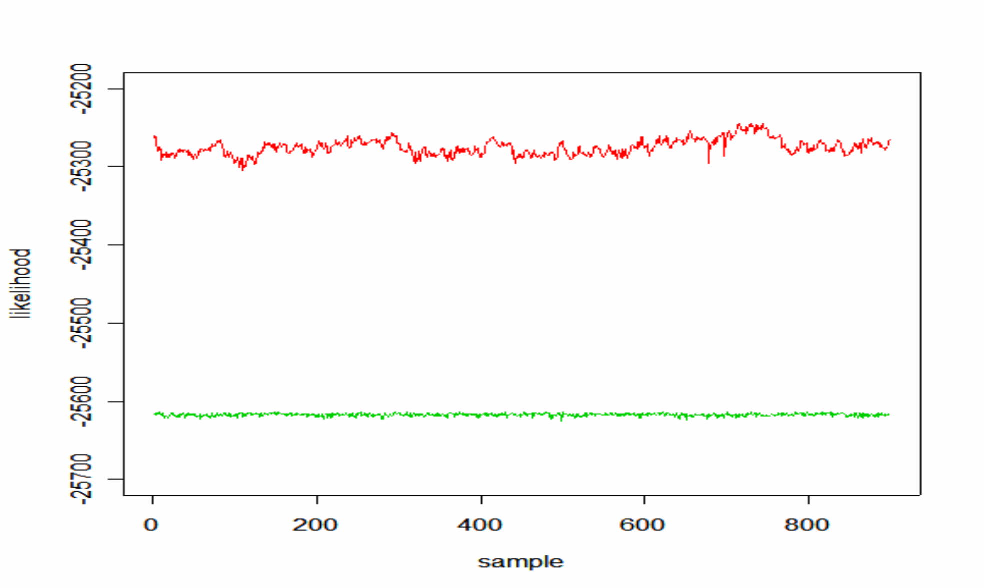

Figure1. The likelihood curves of two analyses for the finch data set. Pink curve is the likelihood for our model. Blue curve is the likelihood for the concatenation method. They are well separated and our model has much greater likelihood than the concatenation method.

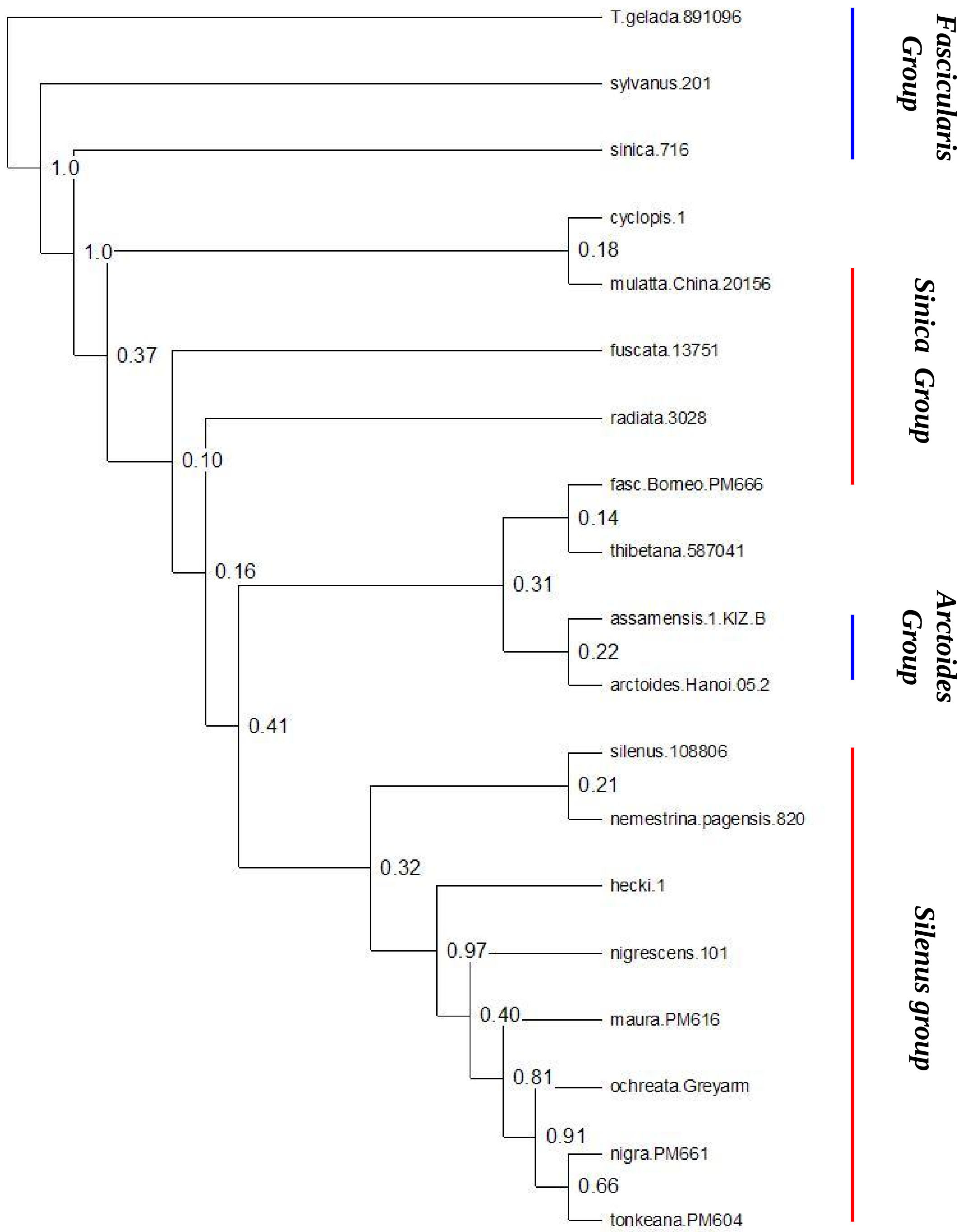

Figure 2. The estimate of species tree for macaques using the BEST method. This is the consensus tree of the sample trees from the posterior distribution of species trees. *T.gelda* is outgroup. The species tree has 4 species groups. The four groups identified by the species tree are strongly correlated with the geographical distribution of the species.