Estimating Species Phylogenies Using Coalescence Times among Sequences

Liang Liu[1,*], Lili Yu[2], Dennis K. Pearl[3], and Scott V. Edwards[1]

[1]*Department of Organismic and Evolutionary Biology, Harvard University, Cambridge, MA 02138, USA; E-mail: lliu@oeb.harvard.edu,*
sedwards@fas.harvard.edu;
[2]*Department of Biostatistics, Georgia Southern University, Statesboro, GA 30460, USA; E-mail: lyu@georgiasouthern.edu; and*
[3]*Department of Statistics,The Ohio State University, Columbus, OH 43210, USA; E-mail: dkp@stat.osu.edu;*
*Correspondence to be sent to: Department of Organismic and Evolutionary Biology, Harvard University, 26 Oxford Street, Cambridge, MA 02138, USA;
E-mail: lliu@oeb.harvard.edu.

Abstract.—The estimation of species trees (phylogenies) is one of the most important problems in evolutionary biology, and recently, there has been greater appreciation of the need to estimate species trees directly rather than using gene trees as a surrogate. A Bayesian method constructed under the multispecies coalescent model can consistently estimate species trees but involves intensive computation, which can hinder its application to the phylogenetic analysis of large-scale genomic data. Many summary statistics–based approaches, such as shallowest coalescences (SC) and Global LAteSt Split (GLASS), have been developed to infer species phylogenies for multilocus data sets. In this paper, we propose 2 methods, species tree estimation using average ranks of coalescences (STAR) and species tree estimation using average coalescence times (STEAC), based on the summary statistics of coalescence times. It can be shown that the 2 methods are statistically consistent under the multispecies coalescent model. STAR uses the ranks of coalescences and is thus resistant to variable substitution rates along the branches in gene trees. A simulation study suggests that STAR consistently outperforms STEAC, SC, and GLASS when the substitution rates among lineages are highly variable. Two real genomic data sets were analyzed by the 2 methods and produced species trees that are consistent with previous results. [Coalescent model; gene tree; species tree.]

Estimating species trees (as distinct from gene trees) is an important but challenging problem in evolutionary biology. For multilocus sequences data, the concatenation (or supermatrix) method, in which all genes are concatenated and analyzed as a single tree of multiple genes, has been commonly used to estimate species trees (Huelsenbeck et al. 1996; William and Ballard 1996). Aside from the fundamental priority of species trees over gene trees as a goal of molecular systematics (Edwards 2009), species tree approaches that allow gene tree heterogeneity have a number of advantages over supermatrix approaches. Recent simulation studies show that the concatenation method may be inconsistent when the species tree is in the anomaly zone (a class of species trees whose most common gene tree is topologically different due to very short branches in the species tree as measured in coalescent units; Degnan and Rosenberg 2006) because the assumption of homogeneous tree topologies across genes on which the concatenation method is based is seriously violated in such cases (Kubatko and Degnan 2007). The concatenation approach has also been shown to be inconsistent for some trees outside but close to the anomaly zone (Edwards et al. 2007). By contrast, a recently proposed Bayesian method (Bayesian estimation of species trees or BEST; Edwards et al. 2007; Liu and Pearl 2007; Liu et al. 2008), which allows heterogeneous gene trees among loci, is able to consistently estimate the species tree even when the species tree is in the anomaly zone (Liu and Edwards 2009). However, the intensive computation behind the Bayesian method hinders its application to large data sets, such as phylogenomic data sets that may contain hundreds of genes. To analyze phylogenomic data sets, methods based onsimple computa-

tion may be more suitable than computationally intensive methods such as BEST, although simpler methods may require more data (genes) in order to achieve a certain level of accuracy in the species tree estimates.

Recent analyses of traditional phylogenetic methods have studied the relationship between gene trees and species trees in the context of coalescent theory (Kingman 1982, 2000). For example, it is now known that simple consensus, which estimate species tree topologies by summarizing gene tree topologies, may result in inconsistent estimates of species trees in the anomaly zone, whereas likelihood supertree methods may be consistent in such situations (Steel and Rodrigo 2008; Degnan et al. 2009). A variety of species tree estimation methods have been shown to be consistent when gene trees are known without error. The rooted triple approach, which combines a set of rooted trees of 3 taxa to produce a species tree, can consistently estimate the species tree (Ewing et al. 2008; Degnan et al. 2009). The probability distribution of the coalescence time in the context of the multispecies coalescent has been studied by several authors (Takahata 1989; Rannala and Yang 2003; Efromovich and Kubatko 2008). Under the coalescent model, Maddison and Knowles (2006) proposed to cluster species by the shallowest coalescences occurring between 2 species based on the fact that minimum coalescence times are consistent estimates of the species divergence times (Takahata 1989). Additionally, Mossel and Roch (2008) have shown that the species tree constructed from minimum coalescence times across genes, that is, the Global LAteSt Split (GLASS) tree (also known as the maximum tree; Liu et al. 2009), is a consistent estimate of the species tree (topology and branch length).

Recent studies have shown that under the multispecies coalescent model (Rannala and Yang 2003), the tree of average coalescence times is a consistent estimate of the species tree topology under a molecular clock (Liu and Edwards 2009). Although these methods estimate the species tree from gene trees assuming gene trees are given and known without error, they can be extended to be employed on actual sequence data using approaches such as the multilocus bootstrap (Seo 2008).

It has been well appreciated that gene trees may be incongruent with the species tree (Pamilo and Nei 1988; Maddison 1997; Rosenberg 2002; Maddison and Knowles 2006; Rosenberg and Tao 2008). The discrepancies between the gene trees and the species tree can be explained by many biological phenomena, such as deep coalescence, horizontal transfer, and gene duplication and loss (Maddison 1997). The methods we develop here are based on the multispecies coalescent model (Rannala and Yang 2003), which assumes that discrepancies between the gene trees and the species tree are exclusively due to deep coalescence, likely among the most common causes of discrepancy (Edwards 2009). It is also assumed that there is no selection. In addition, the population in the species tree is panmictic, and there is no recombination within each gene but free recombination between genes.

EXPECTED RANKS OF COALESCENCES AMONG SEQUENCES

For a rooted gene tree, we define the ranks of coalescences as follows. The rank of the coalescence at the root node is equal to the number of taxa in the tree. The rank decreases by 1 as the node goes from the root toward to the tips of the gene tree (Fig. 1a). Note that all ranks are positive and depend on the topology of the gene tree. We use $r(C_{ab}, g)$ to denote the rank of coalescence C_{ab} of the alleles a and b in gene tree g and $E_g(r(C_{ab}, g)|S)$ to denote the expected rank of coalescence C_{ab} over the gene trees generated from the species tree S with the probability distribution specified by the formula of Rannala and Yang (2003) .

Consider 2 ancestral populations (APs), AP_1 and AP_2, in the species tree S of species A, B, C, and D (Fig. 1b) in which AP_1 is the descendant population of AP_2. Let a, b, c, and d be DNA sequences (alleles) sampled from species A, B, C, and D (Fig. 1b). The most recent common AP of the sequences a and b is AP_1, whereas the most recent common AP of the sequences a and c (or b and c) is AP_2. Let C_{ab} be the coalescence event of a and b and C_{ac} be the coalescence event of a and c in the gene tree g randomly generated from the species tree S using the formula of Rannala and Yang. Sequences a and b either coalesce in population AP_1 or not coalesce in population AP_1. When they coalesce in population AP_1 ($C_{ab} \in AP_1$), the rank of C_{ab} is less than that of C_{ac}, that is, $r(C_{ab}, g) < r(C_{ac}, g)$. Thus, given the species tree S, the expected rank of C_{ab} is less than that of C_{ac} if sequences a and b coalesce in population AP_1,

$$E_g(r(C_{ab}, g)|C_{ab} \in AP_1, \ S) < E_g(r(C_{ac}, g)|S). \quad (1)$$

If sequences a and b do not coalesce in AP_1 ($C_{ab} \notin AP_1$), they will coalesce in one of the populations ancestral to AP_1 under the assumption that genes coalesce before species diverge. Thus, in this case, sequences a, b, and c will enter into population AP_2. Under the multispecies coalescent model, sequences within each AP of the species tree are equally likely to coalesce with each other, and the order of the coalescences among sequences is uniformly distributed (Wakeley 2008). Thus, when $C_{ab} \notin AP_1$, the expected rank of the coalescence between sequences a and b is equal to that of the coalescence between sequences a and c, that is,

$$E_g(r(C_{ab}, g)|C_{ab} \notin AP_1, \ S) = E_g(r(C_{ac}, g)|S). \quad (2)$$

From (1) and (2), we have

$$
\begin{aligned}
E_g(r(C_{ab}, g)|S) &= E_g(r(C_{ab}, g)|C_{ab} \in AP_1, S) \\
&\quad \times P(C_{ab} \in AP_1|S) \\
&\quad + E_g(r(C_{ab}, g)|C_{ab} \notin AP_1, S) \\
&\quad \times P(C_{ab} \notin AP_1|S) \\
&< E_g(r(C_{ac}, g)|S) \times P(C_{ab} \in AP_1|S) \\
&\quad + E_g(r(C_{ac}, g)|S) \times P(C_{ab} \notin AP_1|S) \\
&= E_g(r(C_{ac}, g)|S), \quad (3)
\end{aligned}
$$

which shows that the order of the expected ranks of the coalescences among sequences is consistent with the ancestral order of populations in the species tree. As a result, the topology of the tree of the expected ranks is identical to that of the species tree (Fig. 1b), which is true for any species tree of arbitrary size, including species trees in the anomaly zone.

In general, the tree of expected ranks of the coalescences has the same topology as that of the species tree,

$$Tr_{top} = S_{top}, \quad (4)$$

where Tr_{top} is the topology of the tree of expected ranks and S_{top} is the topology of the species tree. When the expected ranks of coalescences are known, the tree constructed from a distance matrix in which the entries are twice the expected ranks of coalescences is identical in topology to the tree of the expected ranks of coalescences (Fig. 1b). Such a tree could be constructed by any sort of distance method, such as the neighbor-joining (NJ) method (Saitou and Nei 1987a; Nei and Kumar 2000). Let M_{exp} be the distance matrix of the expected ranks of coalescences (Fig. 1c) and M_{ave} be the distance matrix with expected ranks replaced with average ranks $\sum_{i=1}^{N} r(C_{ab}, g_i)/N$ across genes where $r(C_{ab}, g_i)$ is the rank of the coalescence of the alleles a and b in gene tree g_i and N is the number of genes in the data set. When multiple alleles are sampled from each species, the average rank is equal to the average rank across all genes and all pairs of alleles between 2 species, that is,
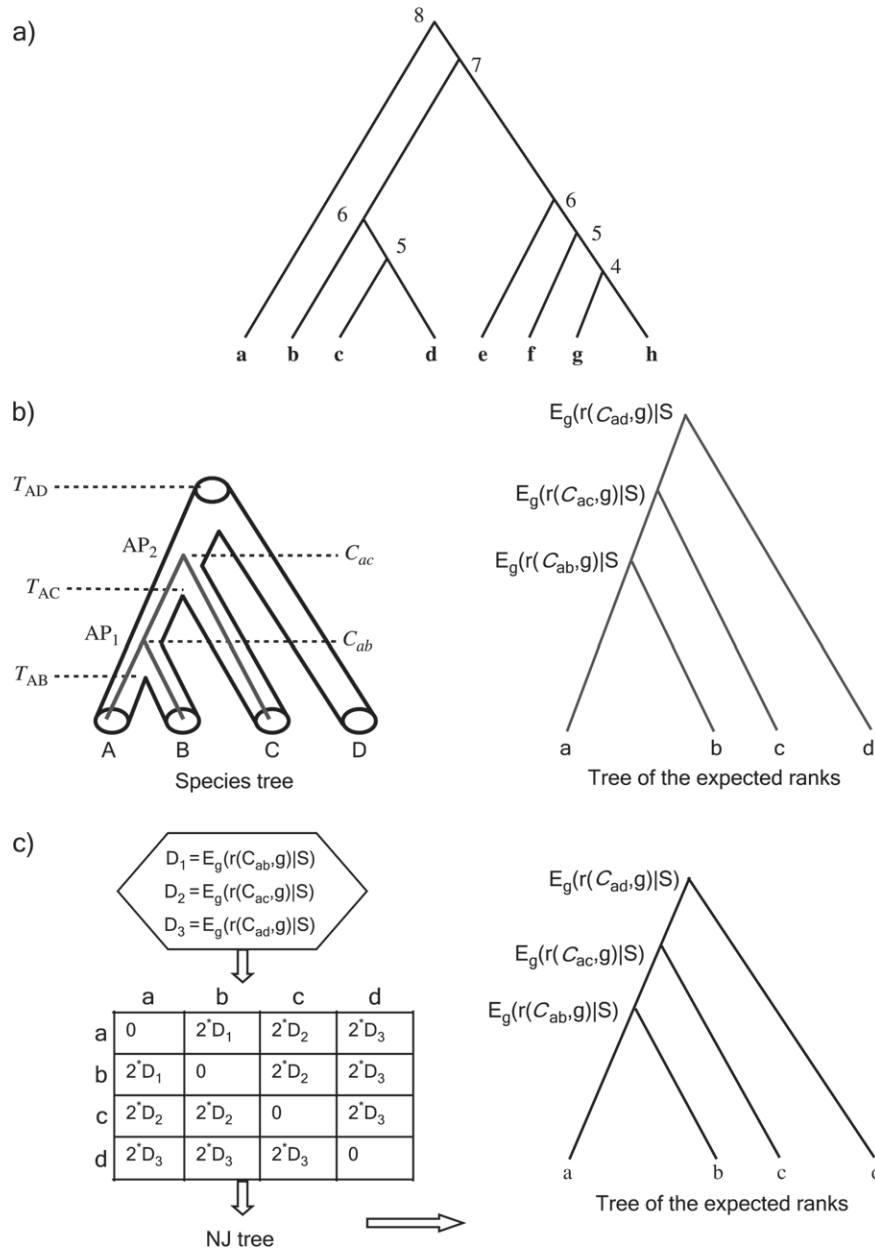
FIGURE 1.　The tree of expected ranks of coalescences among sequences. a) Ranking the coalescences in an 8-taxon gene tree. The coalescence at the root node has rank 8, which is the number of taxa in the gene tree. The rank decreases as it goes from the root to the tips of the gene tree. b) The species tree and the tree of the expected ranks of coalescences for 4 species. The sequences a, b, c, and d are sampled from species A, B, C, and D in the species tree (bold lines). The coalescences (thin lines) for sequences a, b, c, and d occur along the branches of the species tree. We use T_{AB} to denote the species divergence time of species A and B. We use C_{ab} to denote the coalescence of sequences a and b. c) The flow chart for reconstructing the tree of the expected ranks of coalescences. When the expected ranks are given, the NJ tree constructed from the distance matrix in which entries are twice the expected ranks is identical to the tree of the expected ranks of coalescences.

$\frac{1}{K \times J \times N} \sum_{i=1}^{N} \sum_{j=1}^{J} \sum_{k=1}^{K} r(C_{a_j b_k}, g_i)$, where $\{a_1, \ldots, a_J\}$ are J alleles from species A and $\{b_1, \ldots, b_K\}$ are K alleles from species B. By the law of large numbers, the average rank is a consistent estimator of the expected rank of a coalescence event and the distance matrix of average ranks converges to the distance matrix of expected ranks in probability, that is,

$$M_{ave} \xrightarrow{p} M_{exp}, \qquad (5)$$

as the number of genes goes to infinity. Hence, the topology NJ_{top} of the NJ tree constructed from the matrix M_{ave} converges in probability to the topology Tr_{top} of the tree of expected ranks of coalescences,

$$NJ_{top} \xrightarrow{p} Tr_{top}, \qquad (6)$$

as the number of genes goes to infinity. It follows from Equations (4) and (6) that NJ_{top} converges to the species

tree topology S_{top} in probability,

$$\text{NJ}_{\text{top}} \xrightarrow{p} S_{\text{top}}, \qquad (7)$$

as the number of genes goes to infinity. This proves that the NJ tree constructed from the matrix M_{ave} is a consistent estimator of the species tree topology. This same feature holds for other distance methods, such as the Unweighted Pair Group Method with Arithmetic mean or Fitch Margoliash method (Felsenstein 2004). It motivates the method of estimating species trees using average ranks of coalescences among sequences (STAR).

ESTIMATING SPECIES TREES USING AVERAGE RANKS OF COALESCENCE TIMES

The STAR method estimates the species tree by 2 steps. First, a single gene tree is constructed for each locus using any generally consistent method, such as the maximum likelihood (ML) method, without the molecular clock assumption (Felsenstein 1981, 2004). Estimated gene trees are rooted by an outgroup, and the ranks between all pairs of species are counted for each gene tree. In the second step, an NJ tree (or a tree from some other consistent distance method) is constructed from a distance matrix in which the entries are twice the average ranks of coalescences in gene trees across loci. By Equation (7), this NJ tree is a consistent estimate of the species tree topology.

ESTIMATING SPECIES TREES USING AVERAGE COALESCENCE TIMES

Similarly, it can be shown that the tree of the average coalescence times has the same topology as that of the species tree (Liu and Edwards 2009), and the distance tree constructed from a matrix of twice the average coalescence times in gene trees is a consistent estimator of the topology of the species tree when gene trees are given, that is,

$$\text{NJ}_{\text{top}} \xrightarrow{p} S_{\text{top}}, \qquad (8)$$

as the number of genes goes to infinity. This result motivates a method for species tree estimation using average coalescence times (STEAC). However, coalescence times are not given and must be estimated from DNA sequence data. If the substitution rates along the branches in the gene tree are close to each other, that is, the gene tree nearly has a molecular clock, there is correspondence between the order of interspecific coalescences and distances among sequences that are measured as the pairwise distances of taxa in the gene tree (Maddison and Knowles 2006). STEAC consists of 2 consecutive steps. First, gene trees are estimated from multilocus sequences using the ML method without the molecular clock assumption or using some other consistent method for gene tree estimation (Felsenstein 1981, 2005; Saitou and Nei 1987b). Then, the distances among sequences are used to build a STEAC tree as the estimate

of the species tree. Note that both STEAC and STAR can only estimate the topology of the species tree, although STEAC trees can yield somewhat biased species tree branch lengths. An R package implementing STAR and STEAC is available at www.stat.osu.edu/~liuliang (the package will be uploaded to R repository in the future).

INCORPORATING GENE TREE UNCERTAINTY INTO ESTIMATES OF SPECIES TREES

Because gene trees are generally unknown, they must be estimated from sequence data. Two statistical approaches can be used to incorporate uncertainty in the estimated gene trees. We can use a nonparametric bootstrapping technique (Efron 1981) to resample nucleotides within and across genes. Specifically, genes in the multilocus data set are resampled with replacement and then DNA sequences are subsequently resampled with replacement for each gene. (Soltis P.S. and Soltis D.E. 2003; Seo 2008). A consensus tree (Margush and McMorris 1981) can then be constructed from the species trees estimated by the STAR method for the bootstrapped data sets. Alternatively, uncertainty in gene trees can be measured by using the posterior distribution of gene trees estimated by Bayesian phylogenetic approaches. Each sample of the estimated posterior distribution of gene trees is then used to build a STAR tree. These STAR trees can in turn be summarized by a consensus tree. The consensus tree is used as the estimate of the species tree.

COMPARISON OF STAR, STEAC, SHALLOWEST COALESCENCES, AND GLASS

STEAC, shallowest coalescences (SC), and GLASS each differ in the way they summarize coalescence times. For multiple loci, STEAC and SC use average coalescence times across loci, whereas GLASS uses minimum coalescence times across loci, as the distance between 2 species. For multiple alleles, STEAC measures the distance between 2 species as the average of the coalescence times across all pairs of alleles between 2 species, whereas SC and GLASS estimate the distance between 2 species by the minimum coalescence time among all alleles. When extending to sequence data, STEAC, SC, and GLASS have assumptions on the substitution rates so that the distances are consistent with the order of the interspecific coalescences. Serious divergence from the molecular clock may result in systematic errors in estimating the minimum and expected coalescence times and thereby compromise the accuracy of the species tree estimates given by STEAC and SC. Although Mossel and Roch (2008) have shown that GLASS is statistically consistent without the molecular clock assumption, they in fact assume that the substitution rate is the same for all genes and sequences in the same population in the species tree, but the rates may differ across populations. By contrast, STAR uses the rank of coalescence times and does not rely on any assumptions

on the substitution rates, provided that the individual gene trees (topology) are accurately reconstructed.

We used simulation to investigate the effect of divergence from the molecular clock assumption on the performance of STAR, STEAC, SC, and GLASS. There are 2 types of variation among loci and gene trees we can investigate: the variation due to deep coalescence and the variation due to substitution rates. Both of these contribute to estimation errors of these methods. In this simulation study, we also want to know how the 2 types of variation affect the performance of these summary statistics–based methods.

Thirty species trees, each of 20 taxa, were randomly generated from a Yule process implemented in Mesquite (Maddison W.P. and Maddison D.R. 2009). The branches in the generated species trees are in mutation units. Population sizes were randomly generated from the uniform distribution (0.001, 0.01). On average, there are about 20 ($\pm$15) topologically different gene trees per 100 gene trees when averaged across all 30 species trees. Two sequences were sampled from each species except for the outgroup, which had only one sequence. DNA sequences were simulated (assuming Jukes-Canter model) from the gene trees generated from the species trees using the coalescent model in the phylogenetic program MCMCcoal (Rannala and Yang 2003). The generated gene trees were first used as the data to estimate species trees. Because gene trees are given without error, the performance of STAR, STEAC, SC, and GLASS here is solely affected by the interlocus variation due to deep coalescence. Subsequently, the sequence data were used to infer species trees. By comparing the results from sequence data with those from gene trees, we can measure the stochastic effect of mutation on the performance of these species tree estimation approaches.

The model implemented in MCMCcoal to generate sequence data assumes a molecular clock for both species trees and gene trees. The estimation procedure includes 2 steps if DNA sequences are used as data to infer species trees. First, ML gene trees were estimated from each locus using the phylogenetic program PHYML (Guindon and Gascuel 2003). Although the gene trees were simulated assuming a molecular clock, the ML gene trees were estimated without a clock and rooted by an outgroup. The estimated gene trees were then used to reconstruct species trees using STAR, STEAC, SC, and GLASS.

In the second simulation, the assumption of a molecular clock is relaxed. We assume that the substitution rate is the same for all genes and sequences in the same population in the species tree, but the rates may differ across populations (Mossel and Roch 2008). Thus, the gene trees generated from the species tree are not ultrametric. DNA sequences were generated from the non-clocklike gene trees simulated from the species tree (using Phybase, an R package available at www.stat.osu.edu/~liuliang) and used to estimate the species tree. We used the same species trees in the previous simulation, but the terminal and internal branches (terminal or APs) of the species tree were assigned with

relative mutation rates generated from a Dirichlet distribution with all shape parameters equal to each other, that is, $\beta_1 = \beta_2 = \ldots = \beta$. The branches of the gene tree entering a particular population in the species tree are multiplied by the relative mutation rate of that population. If the relative mutation rate is greater than 1, the gene tree branch lengths are increased relative to other gene tree branches. If the relative mutation rate is less than 1, the gene tree branch lengths are decreased. When the relative mutation rates are all equal to 1, the gene trees become clocklike trees. The mean of the relative mutation rates is always equal to 1, whereas the variance is determined by the parameter β of the Dirichlet distribution. We set 3 values for β: 5, 25, and 50. The variances of the relative mutation rates for β= 5, 25, and 50 are 0.199, 0.039, and 0.019, respectively. Small values of β indicate that the variance of the relative mutation rates is large, and the generated relative mutation rates are seriously diverged from their mean 1 (or molecular clock). As in the first simulation, both generated gene trees and DNA sequences were used to estimate species trees using STAR, STEAC, SC, and GLASS.

Overall, the simulation results suggest that the performance of these 4 methods declines when the estimation procedure starts from DNA sequences, especially when there are a small number (10 and 20) of genes. In the first simulation, in which gene trees are given without error and a molecular clock holds, STAR, STEAC, SC, and GLASS perform almost equally well (Fig. 2a). When estimating species trees from DNA sequences, STEAC and SC outperform STAR and GLASS (Fig. 2b). In the second simulation, in which the molecular clock assumption is relaxed, STAR has the best performance consistently for all values of β (Fig. 2c–h), whereas the performance of STEAC, SC, and GLASS declines as the branch lengths deviate from the molecular clock. These results suggest that STAR performs consistently well for the situations that substitution rates are highly variable and where the performance of STEAC, SC, and GLASS is generally poor.

The STAR method assumes that the root of the gene tree is given without error. To investigate the effect of incorrectly rooting gene trees on the performance of STAR, we simulated a 20-taxon species tree in Mesquite (Maddison W.P. and Maddison D.R. 2009). Due to deep coalescence, the gene trees generated from the species tree may have outgroups that are different from that of the species tree. The proportion of gene trees with an outgroup different from the species tree is determined by the internode length between the root and its descendant node in the species tree (the species tree always has a single outgroup taxon). We chose the values 0.001 and 0.1 in substitutions per site for the interbranch length. The corresponding proportions of gene trees with a different outgroup are 0.4 and 0, respectively. Although the outgroups of some simulated gene trees are different from that of the species tree, all gene trees are nonetheless rooted by the outgroup of the species trees in the STAR method to estimate the
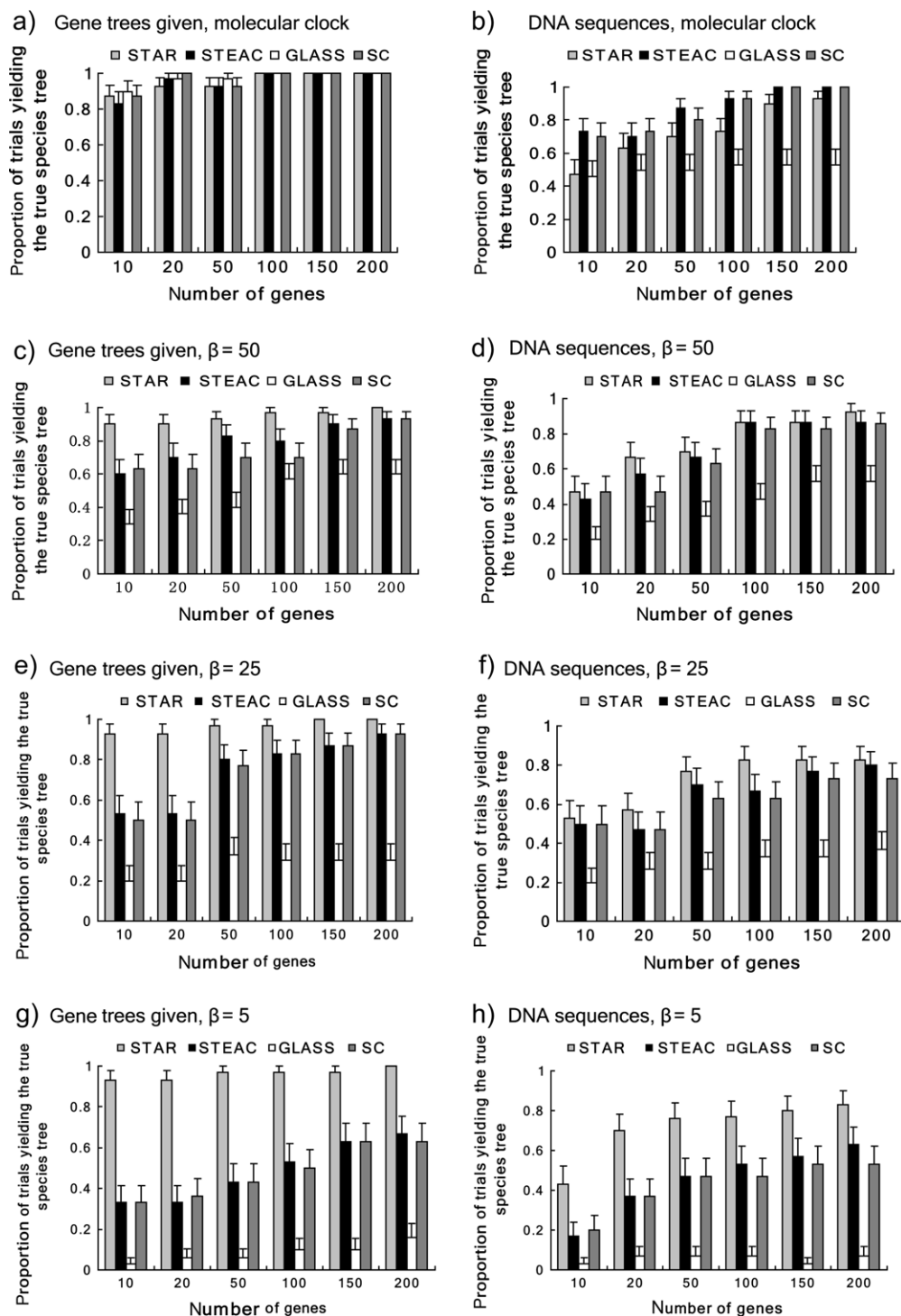
FIGURE 2. Proportion of trials yielding the true species tree for the STAR, STEAC, SC, and GLASS methods. a) Gene trees were generated from 30 clocklike species trees of 20 taxa and used as data to infer species trees using STAR, STEAC, SC, and GLASS. b) DNA sequences were generated from the gene trees simulated in (a) and used to estimate species trees using STAR, STEAC, SC, and GLASS. c) Gene trees were generated from 30 non-clocklike species trees with $\beta = 50$ and used as data to infer species trees. d) DNA sequences were generated from the gene trees simulated in (c) and used as the data to estimate species trees using STAR, STEAC, SC, and GLASS. e) Gene trees were generated from 30 non-clocklike species trees with $\beta = 25$ and used as data to infer species trees. f) DNA sequences were generated from the gene trees simulated in (e) and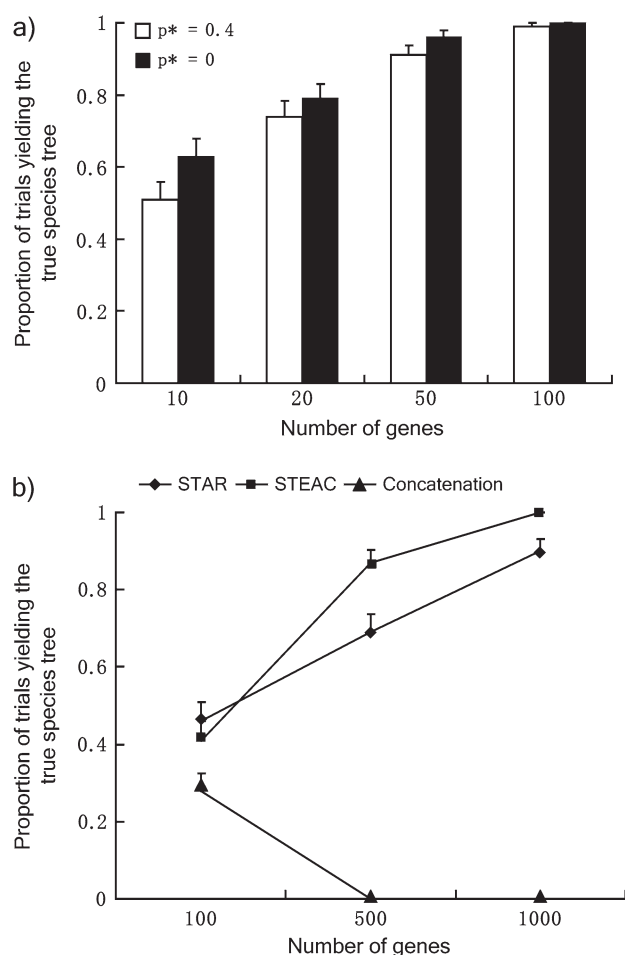 used as the data to estimate species trees. g) Gene trees were generated from 30 non-clocklike species trees with $\beta = 5$ and used as data to infer species trees. h) DNA sequences were generated from the gene trees simulated in (g) and used as the data to estimate species trees.

To investigate the performance of STEAC, STAR in the anomaly zone, we generated DNA sequences from 100, 500, and 1000 gene trees simulated from a species tree: ((((D:0.005, E:0.005):0.00025, C:0.00525):0.00025, B:0.0055):0.01, A:0.0155) with population size $\theta = 0.01$ for all populations. The most probable gene tree generated from this species tree is (((CB)(DE))A), which is different from the species tree, indicating that the species tree is in the anomaly zone. The simulated DNA sequences were analyzed by STEAC, STAR, and the concatenation method to estimate the species tree. For the concatenation method, we used Mrbayes (Huelsenbeck and Ronquist 2001) to reconstruct the species tree from the concatenated sequences, and the species tree was estimated by the consensus tree constructed from the estimated posterior distribution of the species tree. The chain ran for 1,000,000 generations, and the initial 100,000 trees were discarded as a burnin. The simulation was repeated 100 times. Our result shows that the proportion of the STAR and STEAC trees matching the anomalous species tree approaches to 1.0, whereas the proportion of the concatenation trees matching the true species tree goes to 0 (Fig. 3b). This result supports the conclusion that the concatenation method may be statistically inconsistent when the species tree is in the anomaly zone (Kubatko and Degnan 2007; Liu and Edwards 2009). It also suggests that STAR and STEAC can consistently recover anomalous species trees as we showed theoretically in the previous section.

Data Analysis

Yeast Data Analysis

We used a data set consisting of 106 genes totaling over 127,000 bp from 8 species of yeast: *Saccharomyces cerevisiae*, *Saccharomyces paradoxus*, *Saccharomyces bayanus*, *Saccharomyces castellii*, *Saccharomyces kluyveri*, *Saccharomyces mikatae*, *Saccharomyces kudriavzevii*, and *Candida albicans* (Rokas et al. 2003). The species *C. albicans* was used as the outgroup. A consensus tree was constructed from the species tree estimates given by STEAC and STAR for the 100 bootstrap data sets resampled from the original data set using the nonparametric bootstrapping technique described in the previous section. The estimate of the species tree given by STEAC and STAR agrees with previous results using both supermatrix and coalescent approaches (Rokas et al. 2003; Edwards et al. 2007) ((((((*S. paradoxus*, *S. cerevisiae*), *S. mikatae*), *S. kudriavzevii*), *S. bayanus*), *S. castellii*), *S. kluyveri*), with 100% bootstrap support at all nodes. The analysis for 100 bootstrapped samples was performed in parallel on the Odyssey cluster supported by the FAS Research Computation Group. The cluster is build from Dell PowerEdge M6000 with dual Xeon E5410 2.3 Ghz quad core processors and 32 GB RAM. The computation time for each sample was approximately 51 s in which 50 s were taken for PHYML to estimate gene trees and only about 1 s for computing STAR and STEAC trees.

FIGURE 3.   a) The effect of incorrectly rooting gene trees on the performance of STAR. Gene trees were generated from the species tree using the coalescence model and rooted by the outgroup of the species tree. The length of the branch between the root and its descendant node in the species tree changed from 0.1 to 0.001, resulting the proportion p* of gene trees with a different outgroup to increase from 0 to 0.4. The gene trees rooted by the outgroup of the species tree were used to infer the species tree in the STAR method. b) The performance of the STEAC, STAR, and concatenation methods in the anomaly zone. We simulated 500 bp of sequence from a species tree in the anomaly zone: ((((D:0.01, E:0.01):0.0005, C:0.0105):0.0005, B:0.011):0.02, A:0.031) with $\theta = 0.02$ for all populations. The simulated DNA sequences were analyzed by STEAC, STAR, and the concatenation method to estimate the species tree. For the concatenation method, we used Mrbayes (Huelsenbeck and Ronquist 2001) to reconstruct the species tree from the concatenated sequences.

species tree. Our result suggests that the performance of STAR becomes only slightly worse as the proportion of gene trees with a different outgroup increases from 0 to 0.4 (Fig. 3a), indicating that incorrectly rooting gene trees does not have major effects on the performance of STAR. In addition, as the number of genes increases, STAR can consistently recover the true species tree even when 40% of the gene trees are rooted with a wrong outgroup, such as would commonly be encountered, for example, in the anomaly zone.

Mammal Data Analysis

Springer et al. (2007) used mutilocus DNA sequences to reveal the phylogenetic relationship among mammals with *Opossum* as the outgroup. We slightly reduced their data set from 57 to 54 species so that a single sequence could be used to represent the outgroup. The data set we used contains DNA sequences from 20 genes for 54 mammals, totaling 14,326 sites. The phylogenetic analysis for this mammal data set produced highly incongruent gene trees, which implies shallow divergence of these mammals relative to their population sizes. There are 4 major clades: Xenarthra, Laurasiatheria, Euarchontoglires, and Afrotheria in the
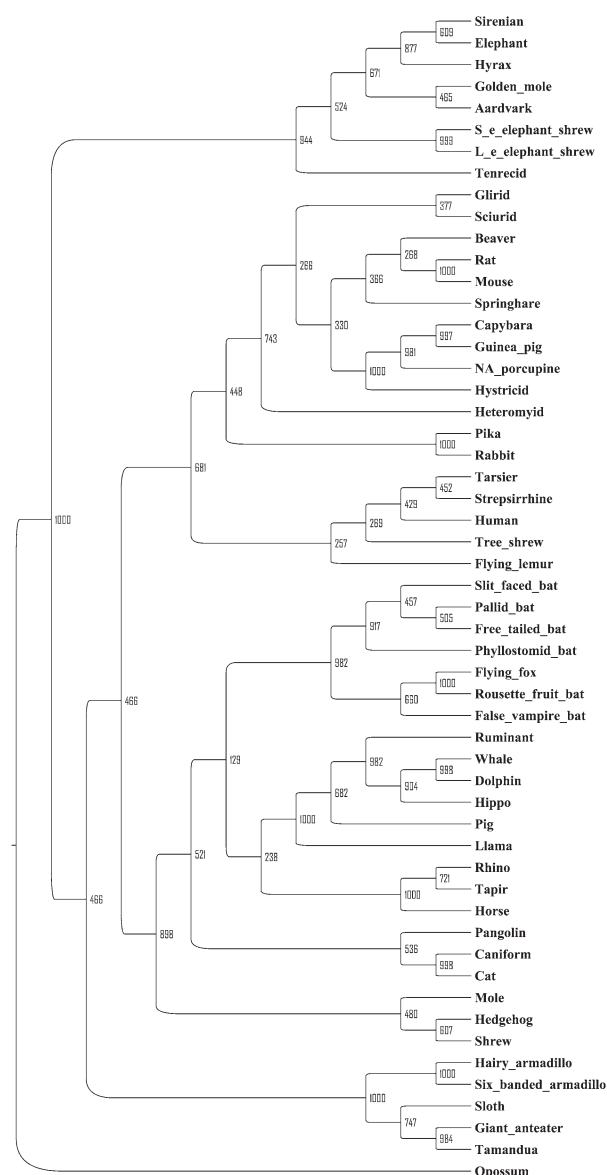

FIGURE 4.   The estimate of the species tree for the mammal data set. The species tree was estimated by a consensus tree constructed from the STAR or STEAC trees for 1000 bootstrapped samples. Both methods produced the same tree. Abbreviations: X = Xenarthra, L = Laurasiatheria, E = Euarchontoglires, and A = Afrotheria.

tree reconstructed by Mrbayes for the concatenated sequences (Springer et al. 2007). In our analysis, the species tree was estimated by STEAC and STAR for each of 1000 bootstrap samples generated from the original data set using a nonparametric bootstrapping technique. Both methods estimated the same species tree (Fig. 4); this species tree is fairly similar to the concatenation result (Springer et al. 2007) but with overall less support. The topological distance (RF distance; Robinson and Foulds 1981) between the STAR/STEAC tree and the concatenation tree is 20, significantly smaller than the distance between the concatenation tree and a random tree. The STAR/STEAC tree identifies exactly the same 4 major groups (Fig. 4) in the concatenation tree (Springer et al. 2007). Additionally, the ancestral relationship among these 4 major groups reconstructed by STAR and STEAC is consistent with previous results (Springer et al. 2007). The STAR analysis was performed on the Harvard University Odyssey cluster. The computation time for each bootstrap sample was approximately 252 s in which 250 s were consumed by PHYML for estimating gene trees and only about 2 s for computing STAR and STEAC trees.

DISCUSSION

Estimating the evolutionary history of species is a central task in evolutionary biology. Multilocus sequence data are useful in the estimation of species phylogenies (Maddison 1997; Edwards and Beerli 2000; Rosenberg 2002; Degnan and Salter 2005; Delsuc et al. 2005). Most current models for multilocus sequence data involve 2 stochastic processes: the coalescent process and the mutation process (Efromovich and Kubatko 2008; Liu et al. 2008). The former is used to model the relationship between the gene trees and the species tree, whereas the latter is used to explain the nucleotide variation observed among sequences. These 2 intimately related processes establish the mathematical foundation for modeling species evolution and phylogenies. Model-based methods such as BEST (Liu and Pearl 2007; Liu et al. 2008) utilize full information of the data to estimate the species tree. Because these methods involve intensive computation, using these methods to infer the species tree for large genomic data sets are beyond current computational resources. Alternatively, methods based on summary statistics such as the STAR and STEAC methods proposed here are based on simple computation and are able to estimate species trees for large-scale genomic data sets rapidly. For example, the computational time for constructing STAR and STEAC trees in the simulation studies conducted in this paper was in seconds, whereas BEST took hours for running for 1 million generations, which still did not guarantee the convergence of the Markov chain (Table 1).

According to coalescent theory, coalescence times in gene trees are consistent estimators of species divergence times and can be used to estimate the species tree (Efromovich and Kubatko 2008; Mossel and Roch

TABLE 1. The computation time for STAR, STEAC, SC, GLASS, and BEST

Number of loci	STAR	STEAC	SC	GLASS	BEST (h/1 million generations)
50	5	27	28	26	2.5
100	8	54	54	52	5.5
150	12	80	81	78	9
200	16	110	110	108	Out of memory

Note: Multilocus DNA sequences generated from the 20-taxon species tree in the first simulation (see text) are used to test the computation time for STAR, STEAC, SC, GLASS, and BEST. The computation time for STAR, STEAC, SC, and GLASS are measured in seconds, except for the BEST method.

2008). Gene coalescence times are generally unknown, but they can be estimated from DNA sequence data. Species tree estimation methods based on gene coalescence times such as BEST must involve estimates of, at the very least, relative substitution rates among loci in order to estimate coalescence times. By contrast, STAR does not need to estimate the gene tree branch lengths accurately because it is based only on the ranks of coalescences.

The consistency of STAR, STEAC, SC, and GLASS is based on the assumption that the incongruence between the gene trees and the species tree is exclusively due to the deep coalescence. If the evolutionary process of DNA sequences involves other biological factors such as gene flow and horizontal gene transfer (HGT), STAR, STEAC, SC, and GLASS may not be able to consistently estimate the true species tree. Gene flow and HGT can result in systematic error in minimum coalescence times and therefore have serious effects on the GLASS method, which estimates species trees by minimum coalescence times. By contrast, STAR, STEAC, and SC are more robust to the gene flow and HGT because they are based on the average coalescence times (or ranks), and a few extremely small coalescence times will not have major impact on the average (Maddison and Knowles 2006). However, if gene flow or HGT is the major source in the evolutionary process of sequences, the average coalescence times may be misleading and STAR, STEAC, and SC may consistently produce the incorrect estimates of species trees.

SUPPLEMENTARY MATERIAL

Supplementary material can be found at http://www.oxfordjournals.org/our_journals/sysbio/.

FUNDING

ACKNOWLEDGEMENTS

REFERENCES

Degnan J.H., DeGiorgio M., Bryant D., Rosenberg N.A. 2008. Properties of consensus methods for inferring species trees from gene trees. Syst. Biol. 58:35–54.

Degnan J.H., Rosenberg N.A. 2006. Discordance of species trees with their most likely gene trees. PLoS Genet. 2:762–768.

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

Delsuc F., Brinkmann H., Philippe H. 2005. Phylogenomics and the reconstruction of the tree of life. Nat. Rev. Genet. 6:361–375.

Edwards S.V. 2009. Is a new and general theory of molecular systematics emerging? Evolution. 63:1–19.

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

Edwards S.V., Liu L., Pearl D.K. 2007. High-resolution species trees without concatenation. Proc. Natl. Acad. Sci. USA. 104:5936–5941.

Efromovich S., Kubatko L.S. 2008. Coalescent time distributions in trees of arbitrary size. Stat. Appl. Genet. Mol. Biol. 7:Article 2.

Efron B. 1981. Nonparametric estimates of standard error—the jackknife, the bootstrap and other methods. Biometrika. 68:589–599.

Ewing G., Ebersberger I., Schmidt H., von Haeseler A. 2008. Rooted triple consensus and anomalous gene trees. BMC Evol. Biol. 8:118.

Felsenstein J. 1981. Evolutionary trees from DNA sequences: a maximum likelihood approach. J. Mol. Evol. 17:368–376.

Felsenstein J. 2004. Inferring phylogenies. Sunderland (MA): Sinauer Associates.

Felsenstein J. 2005. PHYLIP. Seattle (WA): Department of Genome Science, University of Washington.

Guindon S., Gascuel O. 2003. A simple, fast, and accurate algorithm to estimate large phylogenies by maximum likelihood. Syst. Biol. 52:696–704.

Huelsenbeck J.P., Bull J.J., Cunningham C.W. 1996. Combining data in phylogenetic analysis. Trends Ecol. Evol. 11:152–158.

Huelsenbeck J.P., Ronquist F. 2001. Mrbayes: Bayesian inference of phylogenetic trees. Bioinformatics. 17:754–755.

Kingman J.F.C. 1982. On the genealogy of large populations. Stoch. Proc. Appl. 13:235–248.

Kingman J.F.C. 2000. Origins of the coalescent: 1974–1982. Genetics. 156:1461–1463.

Kubatko L.S., Degnan J.H. 2007. Inconsistency of phylogenetic estimates from concatenated data under coalescence. Syst. Biol. 56:17–24.

Liu L., Edwards S.V. 2009. Phylogenetic analysis in the anomaly zone. Syst. Biol. doi: 10.1093/sysbio/syp034.

Liu L., Pearl D.K. 2007. Species trees from gene trees: Reconstructing Bayesian posterior distributions of a species phylogeny using estimated gene tree distributions. Syst. Biol. 56:504–514.

Liu L., Pearl D.K., Brumfield R.T., Edwards S.V. 2008. Estimating species trees using multiple-allele DNA sequence data. Evolution. 62:2080–2091.

Liu L., Yu L., Pearl D.K. 2009. Maximum tree: a consistent estimator of the species tree. J. Math. Biol. doi: 10.1007/s00285-009-0260-0.

Maddison W.P. 1997. Gene trees in species trees. Syst. Biol. 46:523–536.

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

Maddison W.P., Maddison D.R. 2009. Mesquite: a modular system for evolutionary analysis. Version 2.6. Available from: http://mesquiteproject.org.

Margush T., McMorris F.R. 1981. Consensus n-trees. Bull. Math. Biol. 43:239–244.

Mossel E., Roch S. 2008. Incomplete lineage sorting: consistent phylogeny estimation from multiple loci. IEEE/ACM Transactions on Computational biology and Bioinformatics. Available from: http://arxiv.org/abs/0710.0262.

Nei M., Kumar S. 2000. Molecular evolution and phylogenetics. New York: Oxford University Press.

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

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

Robinson D.R., Foulds L.R. 1981. Comparison of phylogenetic trees. Math. Biosci. 53: 131–147.

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

Rosenberg N.A. 2002. The probability of topological concordance of gene trees and species trees. Theor. Popul. Biol. 61:225–247.

Rosenberg N.A., Tao R. 2008. Discordance of species trees with their most likely gene trees: The case of five taxa. Syst. Biol. 57: 131–140.

Saitou N., Nei M. 1987a. The neighbor-joining method—a new method for reconstructing phylogenetic trees. Mol. Biol. Evol. 4: 406–425.

Saitou N., Nei M. 1987b. On the maximum-likelihood method for molecular phylogeny. Jpn. J. Genet. 62:547–548.

Seo T.-K. 2008. Calculating bootstrap probabilities of phylogeny using multilocus sequence data. Mol. Biol. Evol. 25:960–971.

Soltis P.S., Soltis D.E. 2003. Applying the bootstrap in phylogeny reconstruction. Stat. Sci. 18:256–267.

Springer M.S., Burk-Herrick A., Meredith R., Eizirik E., Teeling E., O'Brien S.J., Murphy W.J. 2007. The adequacy of morphology for reconstructing the early history of placental mammals. Syst. Biol. 56:673–684.

Steel M., Rodrigo A. 2008. Maximum likelihood supertrees. Syst. Biol. 57:243–250.

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

Wakeley J. 2008. Coalescent theory: an introduction. Greenwood Village (CO): Roberts & Company Publishers.

William J., Ballard O. 1996. Combining data in phylogenetic analysis. Trends Ecol. Evol. 11:334.