Molecular Phylogenetics and Evolution 53 (2009) 320–328

Contents lists available at ScienceDirect

Molecular Phylogenetics and Evolution

journal homepage: www.elsevier.com/locate/ympev

Review

Coalescent methods for estimating phylogenetic trees

Liang Liu [a], Lili Yu [b], Laura Kubatko [c], Dennis K. Pearl [c], Scott V. Edwards [a,*]

[a] Department of Organismic and Evolutionary Biology, Museum of Comparative Zoology, Harvard University, 26 Oxford street, Cambridge, MA 02138, USA
[b] Department of Biostatistics, Georgia Southern University, USA
[c] Department of Statistics, The Ohio State University, USA

ARTICLE INFO

Article history:
Received 14 November 2008
Revised 21 May 2009
Accepted 28 May 2009
Available online 6 June 2009

Keywords:
Coalescent model
Molecular clock
Phylogenomics
Bayesian statistics
Maximum likelihood

ABSTRACT

We review recent models to estimate phylogenetic trees under the multispecies coalescent. Although the distinction between gene trees and species trees has come to the fore of phylogenetics, only recently have methods been developed that explicitly estimate species trees. Of the several factors that can cause gene tree heterogeneity and discordance with the species tree, deep coalescence due to random genetic drift in branches of the species tree has been modeled most thoroughly. Bayesian approaches to estimating species trees utilizes two likelihood functions, one of which has been widely used in traditional phylogenetics and involves the model of nucleotide substitution, and the second of which is less familiar to phylogeneticists and involves the probability distribution of gene trees given a species tree. Other recent parametric and nonparametric methods for estimating species trees involve parsimony criteria, summary statistics, supertree and consensus methods. Species tree approaches are an appropriate goal for systematics, appear to work well in some cases where concatenation can be misleading, and suggest that sampling many independent loci will be paramount. Such methods can also be challenging to implement because of the complexity of the models and computational time. In addition, further elaboration of the simplest of coalescent models will be required to incorporate commonly known issues such as deviation from the molecular clock, gene flow and other genetic forces.

1. Introduction

Phylogeny is used to represent the evolutionary history of species observed through time, and is thus one of the most important entities in evolutionary biology (Hillis et al., 1993; Swofford et al., 1996; Avise, 2000; Ma et al., 2000). It assumes that all species arise from a common ancestor and that genetic material is transmitted from ancestors to descendants along the branches of the phylogenetic tree. Phylogenetic information is encoded in the genetic material of contemporary species in a manner that allows the information from data such as DNA sequences to be used to trace the history back to the most recent common ancestor of the species. While the phylogenetic tree relating the sequences at a single locus, known as a gene tree, has been rigorously studied for decades, research on the phylogeny of species—the entity that contains the genetic variation and is arguably the true focus of phylogenetics—is, ironically, still in its infancy (Liu et al., 2008; Edwards, 2009). The observation of a tremendous amount of variation in gene trees (both in topologies and branch lengths) estimated from multilocus sequence data has stimulated research on the estimation of species-level phylogenies in contexts in which variation at the level of individual genes is taken into account (Pamilo and Nei, 1988; Powell, 1991; Doyle, 1992; Hudson, 1992; Brower et al., 1996; Page and Charleston, 1997; Cao et al., 1998; Pollard et al., 2006). The emphasis of phylogeny is the evolutionary history at the level of species, which may be distinct from the genealogical pathway of individuals, or gene trees (Pamilo and Nei, 1988; Powell, 1991; Nichols, 2001; Rannala and Yang, 2008). The fact that a gene tree is the evolutionary history of alleles randomly chosen from species provides, from a biological perspective, a reasonable explanation for the relationship between gene trees and the phylogeny of species (Pamilo and Nei, 1988; Maddison, 1997). It indicates that a gene tree is a random tree generated within the phylogeny of species and phylogenies of species should be studied in the framework of probabilistic models that incorporate the probability distribution of gene trees given the phylogeny of species. Although a few techniques have been developed to specify this probability distribution in the context of a variety of biological phenomena such as horizontal gene transfer (HGT) and gene duplication/loss (Arvestad et al., 2003; Linz et al., 2007), this review will focus on approaches that assume that the conflicts between gene trees and the species tree are exclusively due to deep coalescence (Maddison, 1997; Maddison and Knowles, 2006). Some promising methods for inferring phylogenies in the presence of horizontal gene transfer have been developed, although these methods do not acknowledge the possibility of gene tree discordance due to deep coalescence (Linz et al., 2007).

* Corresponding author. Fax: +1 617 495 5667.
E-mail address: sedwards@fas.harvard.edu (S.V. Edwards).

1055-7903/$ - see front matter © 2009 Published by Elsevier Inc.
doi:10.1016/j.ympev.2009.05.033

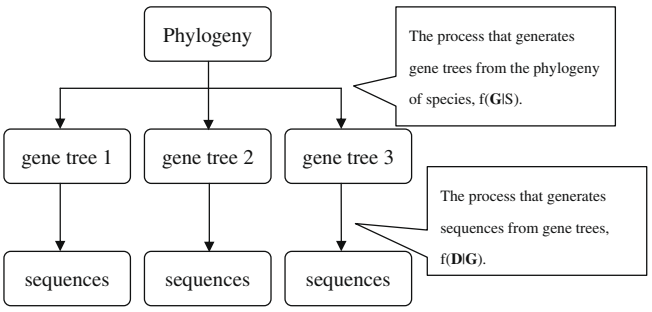


Fig. 1. Flow chart for demonstrating the statistical model for multilocus sequences generated from the phylogeny of species. The chart illustrates the independence of the two major stochastic processes in generating molecular data from species trees: the generation of gene trees from the species tree and the generation of DNA sequences from the constituent gene trees.

The probability distribution of gene trees (**G**) given the species tree, (S), or $f(\mathbf{G}|S)$, can be used to infer species phylogenies when gene trees **G** are known. However, gene trees are generally unknown and phylogenies must be estimated from multilocus data. A probabilistic model for estimating species phylogenies consists of three components; multilocus data (**D**), gene trees (**G**), and the phylogeny of species (S) (Liu and Pearl, 2007). The data will most commonly be nucleotide sequence or amino acid data, but any data that contain phylogenetic information for a gene may be used. The data for a particular gene are the result of a process of descent with modification along the branches of the gene tree, while the gene tree itself is a random tree sampled from a probability distribution dependent on the phylogeny of species (Fig. 1) (Liu et al., 2008). Although gene trees are random conditional on the species tree, this need not mean that gene trees are heterogeneous in topology; depending on the species tree, random gene trees may be highly constrained and thus highly uniform in topology and branch lengths. The sequences are generated from the phylogeny S through two random processes; the process that generates gene trees from the phylogeny of species, which has probability distribution $f(\mathbf{G}|S)$, and the process that generates data from gene trees, which has probability distribution $f(\mathbf{D}|\mathbf{G})$. These two consecutive processes must be simultaneously considered in the development of probabilistic models for genetic data obtained from the phylogeny of species (Fig. 1).

2. The coalescence model for multilocus sequences

We now explicitly consider nucleotide and amino acid sequence data, for which the mutation process describes how the nucleotides in the sequences change through time along the branches of gene trees. For multilocus data, we use D_i and G_i to denote the aligned nucleotides or amino acids of all individuals sampled from the species under study and the gene tree (topology and branch lengths) for locus i, respectively. The probability distribution $f(D_i|G_i)$ of the alignment D_i given the gene tree G_i is the likelihood function traditionally used in the maximum likelihood method for estimating gene trees (Jukes and Cantor, 1969; Felsenstein, 1981; Hasegawa et al., 1985; Whelan and Goldman, 2001; Sullivan, 2005). Assuming independence among loci in a multilocus data set, the likelihood function $f(\mathbf{D}|\mathbf{G})$ is the product of functions $f(D_i|G_i)$ across loci, which is used to measure the fit of gene trees to the multilocus sequence data.

The function $f(\mathbf{G}|S)$ is the probability distribution of gene trees given the phylogeny S. The most commonly considered biological phenomena that contribute to the conflicts of gene trees and the phylogeny of species include deep coalescence, horizontal transfer, and gene duplication/gene loss (Maddison, 1997; Pollard et al., 2006). While the probability distribution $f(\mathbf{G}|S)$ may be derived from any of these biological events either individually or in combination (Eulenstein et al., 1998; Stege, 1999; Huson et al., 2005; Sanderson and McMahon, 2007; Holland et al., 2008), the coalescent is the most studied process in modeling the variation among gene trees due to its mathematical simplicity and the fact that the phylogeny of species consists of multiple ancestral and contemporary populations within which the ancestral history of individuals is commonly modeled by a coalescent process in a population genetics framework under some general conditions (an example of such conditions are listed below). Thus we concentrate on a probabilistic model, called the multispecies coalescent model, derived from the coalescence process (Kingman, 1982; Kingman, 2000; Degnan and Salter, 2005; Degnan et al., 2008; Wakeley, 2008), which assumes that the effect of biological phenomena such as horizontal transfer and gene duplication/gene loss are negligible compared to the effect of coalescence in the evolutionary process of individuals sampled from species. In principle, the multispecies coalescent model (Rannala and Yang, 2003; Degnan and Salter, 2005; Degnan et al., 2008) can be extended to accommodate horizontal transfer and gene duplication/gene loss in order to analyze the datasets in which these biological events are commonly involved, but this has not yet been accomplished.

The model developed for multilocus DNA sequences in the context of the coalescent includes two probability distributions: the probability distribution $f(\mathbf{D}|\mathbf{G})$, which is the likelihood function used for estimating gene trees, and the probability distribution $f(\mathbf{G}|S)$, derived from the multispecies coalescent model (Maddison, 1997; Felsenstein, 2004). The coalescence model assumes that (1) any incongruence between gene trees and the phylogeny of species is exclusively due to the coalescent. However the model does not necessarily assume that there is any incongruence; sets of topologically congruent gene trees, as one would detect in analyses involving species trees with long branches and large internodes, can be analyzed as well. The model also assumes that (2) there is free recombination between genes and no recombination within each gene, (3) there is random mating in each population (current and ancestral) in the phylogeny of species, (4) there is no selection, (5) the mutation process along the lineages in the gene tree follows an evolutionary model (Jukes-Cantor, HKY, or GTR model), (6) sequences at a single locus are conditionally independent of the phylogeny of species if the gene tree for that locus is given, and (7) sequences from different genes are mutually independent if their gene trees are given and the gene trees are mutually independent if the species tree is given. An explicit mathematical formulation of the model described above for multilocus sequences **D**, gene trees **G**, and the phylogeny of species S, then, is as follows:

$$f(\mathbf{D}|\mathbf{G}) = \prod_{i=1}^{K} f(D_i|G_i)$$

$$f(\mathbf{G}|S) = \prod_{i=1}^{K} f(G_i|S),$$

$$f(D_i|G_i, S) = f(D_i|G_i) \quad \text{for } i = 1, \ldots K,$$

where K is the number of genes. The last equation indicates that sequences D_i are conditionally independent of the phylogeny S when the gene tree G_i is given. The function $f(D_i|G_i)$ is the likelihood function derived from nucleotide substitution models (Jukes and Cantor, 1969; Felsenstein, 1981, 2004; Hasegawa et al., 1985), while $f(G_i|S)$ is Rannala and Yang's gene tree density given a species tree (Rannala and Yang, 2003). Let t_{ij} be the time interval between the (j-1)th and jth coalescence in population i. Note that t_{i1} is the time interval between the first coalescence and the species divergence time for population i. The probability density of a gene tree topology and the $(m_i - n_i)$ time intervals $t_{i(n_i+1)}, \ldots, t_{im_i}$ for population i with

effective population size θ_i reduced from m_i to n_i sampled alleles along a branch of length τ_i in a species tree is

$$\exp\left(-\frac{n_i(n_i-1)}{\theta_i}\left(\tau_i-\sum_{j=n_i+1}^{m_i}t_{ij}\right)\right)\prod_{j=n_i+1}^{m_i}\left[\frac{2}{\theta_i}\exp\left(-\frac{j(j-1)}{\theta_i}t_{ij}\right)\right]. \quad (1)$$

The probability distribution of a gene tree **G** (topology and branch length) given the phylogeny S is the product of the probability distribution in (1) across current and ancestral populations (branches) in the phylogeny S,

$$f(G|S)=\prod_{i=1}^{M}\left\{\exp\left(-\frac{n_i(n_i-1)}{\theta_i}\left(\tau_i-\sum_{j=n_i+1}^{m_i}t_{ij}\right)\right)\times\prod_{j=n_i+1}^{m_i}\left[\frac{2}{\theta_i}\exp\left(-\frac{j(j-1)}{\theta_i}t_{ij}\right)\right]\right\}, \quad (2)$$

where M is the number of branches in the phylogeny S.

As mentioned before, the likelihood $f(\mathbf{D}|\mathbf{G})$ is more familiar to phylogeneticists than is the second likelihood $f(\mathbf{G}|S)$. To illustrate the second likelihood function computing the probability density of a gene tree given a species tree, we derive the probability densities of various gene trees given a single species tree and illustrate the dependence of this probability density on population sizes of the species tree. The probability densities of the gene trees embedded in the species phylogeny in Fig. 2 are

$$\frac{2}{\theta_1}e^{-2*(3.5-\tau_1)/\theta_1}*\frac{2}{\theta_2}e^{-2*(2.5-\tau_2)/\theta_2}*\frac{2}{\theta_3}e^{-2*(1.5-\tau_3)/\theta_3}. \quad (3)$$

It follows from (3) that the probability density of a gene tree decreases for large differences between coalescence times in the gene tree and species divergence times τ in the species phylogeny. In Fig. 2a and b we calculate the probability densities of two gene trees given a species tree. The probability density of the gene trees vary as we change the values of the population sizes and divergence times in the species tree. The coalescence times in gene tree 1 of Fig. 2a are closer to the divergence times in the species tree than those in gene tree 2 of Fig. 2b. Correspondingly, the probability densities for gene tree 1 with respect to different values of parameters in the species tree are larger than those for gene tree 2 (Fig. 2c), illustrating the fact that the gene trees topologically concordant with the species tree are usually more probable than the gene trees discordant with the species tree. A notable exception to this pattern occurs when species trees are in the so-called anomaly zone, a region of species tree space – usually with very short internal branches – where discordant gene trees are actually more common (and therefore more likely) than the topologically concordant gene tree (Degnan and Rosenberg, 2006).

The likelihood of the phylogeny S is

$$f(\mathbf{D}|S)=\int_{\mathbf{G}}f(\mathbf{D}|\mathbf{G})*f(\mathbf{G}|S)d\mathbf{G} \quad (4)$$

where the integral is over all possible gene genealogies, including both topologies and branch lengths. As in the likelihood analysis for gene tree estimation, one must evaluate a large number of phylogenetic trees, in this case a large number of species trees, in order to find the maximum likelihood estimate of the phylogeny. But in addition to having to evaluate a large number of species trees, the large number of gene trees for any given species tree means that the above likelihood is impractical to calculate directly for all but the smallest species trees. Even this sobering conclusion does not quite hold if one has sampled many alleles per species, which necessarily vastly increases the number of gene trees to be evaluated, as in traditional phylogenetic analysis when one has sampled a large number of species (Felsenstein, 1988).

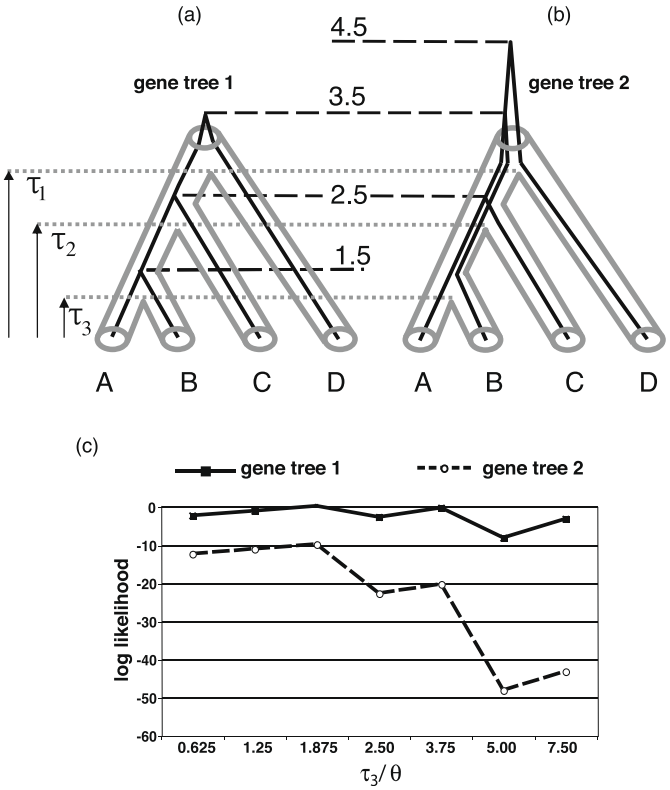


Fig. 2. The probability densities of two gene trees given the phylogeny of four species using the formulation of Rannala and Yang (2003). (a,b) The tree with volume is the phylogeny of species A, B, C, and D; it is identical in both panels. The divergence times for the three populations in the species phylogeny are $\tau_1 = 3.0$, $\tau_2 = 2.0$, and $\tau_3 = 1.0$ and are indicated by gray dotted lines. The ancestral population sizes (not shown) are θ_1, θ_2, and θ_3 at times τ_1, τ_2 and τ_3, respectively. The values of τ and the θ are in arbitrary units but their ratio governs the rate of coalescence (Eq. (3)). For simplicity, we assume that three ancestral populations have the same size, i.e., $\theta_1 = \theta_2 = \theta_3 = \theta$ at all nodes in the species tree, even as the value of θ changes. τ_3 also changes in the species phylogeny but the other τ's do not. The coalescence times of the gene trees embedded in the species phylogenies are 1.5, 2.5, and 3.5 in (a) and 2.5, 3.5, and 4.5 in (b), in arbitrary units and are indicated by black dotted lines. The second gene tree (b) does not match the species phylogeny. (c) The chart indicates the log probability densities of the two gene trees (gene tree 1 and gene tree 2 in (a) and (b), respectively) with respect to different values of ratio τ_3/θ (x axis), given the species tree in (a) and (b).

3. Estimating phylogenies of species: likelihood, Bayesian and summary statistic methods

The phylogeny S can be estimated from multilocus sequence data **D** using the likelihood function $f(\mathbf{D}|S)$. For example, the maximum likelihood estimate (MLE) of the phylogeny S is given by

$$\hat{S}=\arg\max_{S}\{f(\mathbf{D}|S)\}. \quad (5)$$

Bayesian approaches assume a prior distribution for the phylogeny S and use the posterior distribution – the combination of likelihood and prior distributions – to infer phylogenies. The posterior distribution of the phylogeny S is

$$f(S|\mathbf{D})=\frac{f(\mathbf{D}|S)*f(S)}{f(\mathbf{D})}. \quad (6)$$

Unlike the maximum likelihood and Bayesian approaches, which utilize the full data **D** and the likelihood function $f(\mathbf{D}|S)$ to infer the phylogeny of species (as well as prior distributions in the case of Bayesian methods), methods based on summary statistics seek to estimate the phylogeny S by summarizing the gene trees estimated from multilocus sequences. If $\hat{\mathbf{G}}$ is a sufficient statistic (Fish-

er, 1922) of gene trees **G**, it follows from the Fisher's factorization theorem (Casella and Berger, 2002) that the probability density function $f(\mathbf{D}|\mathbf{G})$ is a product of two terms, $f(\mathrm{D}|\mathrm{G}) = k(\mathbf{D}) * f(\hat{\mathbf{G}}|\mathbf{G})$. Then Eq. (4) becomes

$$f(\mathbf{D}|S) = \int_{\mathbf{G}} k(\mathbf{D}) * f(\hat{\mathbf{G}}|\mathbf{G}) * f(\mathbf{G}|S) d\mathbf{G} = k(\mathbf{D}) * f(\hat{\mathbf{G}}|S), \quad (7)$$

which by Fisher's factorization theorem shows that a sufficient statistic for the phylogeny S is $\hat{\mathbf{G}}$. Eq. (7) conveys the important message that phylogenies can be estimated by summarizing the gene trees ($\hat{\mathbf{G}}$) estimated from multilocus sequences. Because it is difficult to derive an analytical expression for a sufficient statistic for the phylogeny S, this statistic may be replaced by other commonly used estimates of gene trees, for example, the MLEs of gene trees (Felsenstein, 1981; Seo et al., 2005). To summarize, methods for estimating species trees can be broadly classified into two general categories: phylogenies estimated using the full data and likelihood function $f(\mathbf{D}|S)$ and phylogenies estimated by summarizing gene trees estimated from multilocus sequences. We discuss each of these general frameworks in more detail below.

3.1. Estimating phylogenies using the full data and likelihood $f(\mathbf{D}|S)$

In general, there are two standard statistical treatments for estimating parameters using likelihood functions, maximum likelihood methods and Bayesian methods. Maximum likelihood approaches estimate the phylogeny S by maximizing the likelihood function $f(\mathbf{D}|S)$. Due to the complexity of the likelihood function, the maximum likelihood estimates are obtained by numerical methods which often involve calculating the likelihood score for an individual phylogeny and updating the phylogeny by modifications of it that have higher likelihood until no further improvement in likelihood scores is observed. The calculation of the likelihood for a phylogeny S is by no means straightforward because the likelihood function $f(\mathbf{D}|S)$ involves an integral over gene trees (see Eq. (4)). For this reason, the maximum likelihood method involves intensive computation for which there is currently no efficient implementation.

Bayesian methods estimate species trees based on the posterior distribution $f(S|\mathbf{D})$. Since the denominator $f(\mathbf{D})$ in the posterior distribution of S in (6) is infeasible to compute, the posterior distribution $f(S|\mathbf{D})$ is approximated by numerical methods such as Markov Chain Monte Carlo (Hastings, 1970) algorithms which also need intensive computation but in general are deemed faster than their likelihood counterparts.

The computational burden can be reduced by simplifying the probability function $f(\mathbf{D}|S)$. For example, the concatenation method (Huelsenbeck et al., 1996; Adachi et al., 2000) assumes homogeneous gene trees across genes and that gene trees are identical to the phylogeny of species, i.e., $G_1 = G_2 \ldots\ldots = G_k = S$, which simplify the function $f(\mathbf{D}|S)$ as

$$f(\mathbf{D}|S) = \int_{\mathbf{G}} f(\mathbf{D}|\mathbf{G}) * f(\mathbf{G}|S) d\mathbf{G} = f(\mathbf{D}|G_1).$$

Thus, the concatenation method employs the likelihood of the gene tree, $f(\mathbf{D}|G_1)$, to estimate the phylogeny S assuming that the gene tree is identical to the phylogeny S. If the assumption of homogeneous gene trees is seriously violated, the concatenation method may be inconsistent in estimating the phylogeny of species (Kubatko and Degnan, 2007). In the case of maximum likelihood, simulation studies suggest that the concatenation method may produce spuriously high bootstrap support for incorrect partitions (Gadagkar et al., 2005; Kubatko and Degnan, 2007). For example, Nishihara et al. (2007) demonstrated that a concatenated analysis of a genomic-scale mammalian data set strongly supported a wrong species phylogeny.

3.2. Estimating phylogenies by summarizing estimated gene trees

Methods in this category estimate species phylogenies by summarizing the gene trees estimated from multilocus sequences (Baum, 1992), using various forms of summarization. In general, these methods can be classified into two groups; nonparametric methods and parametric methods. Nonparametric methods assume no specific distribution for gene trees, while parametric methods typically assume that the probability distribution of gene trees complies with coalescent theory.

3.2.1. Nonparametric methods

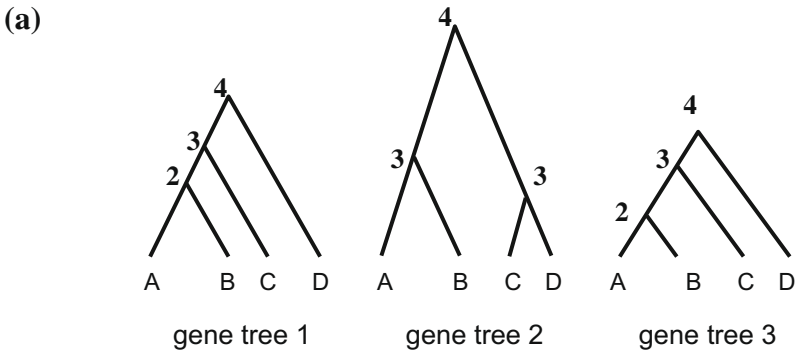
Nonparametric methods include the consensus (Margush and McMorris, 1981; Degnan et al., 2008; Ewing et al., 2008), reconciliation (Page and Charleston, 1997; Slowinski et al., 1997; Page, 1998; Avise, 2000; Page, 2000; V'Yugin et al., 2002; Bonizzoni et al., 2003; Gorbunov and Lyubetsky, 2005; Berglund-Sonnhammer et al., 2006), and supertree (Wilkinson et al., 2005; Cotton and Wilkinson, 2007; Steel and Rodrigo, 2008) methods. Consensus methods use various techniques to construct a single summary tree from the estimated gene trees that is then taken to be the estimated phylogeny of the species. Consensus methods are nonparametric in the sense that they do not assume a specific underlying distribution for the gene trees. The reconciliation method summarizes gene trees by a single tree (or several trees with the same score) that minimizes the number of coalescence, horizontal, and gene duplication/gene loss events required to reconcile gene trees and the query tree (Berglund-Sonnhammer et al., 2006). The supertree method is a family of approaches that merge (or summarize) multiple gene trees into a single tree called the "supertree" (Bininda-Emonds and Bryant, 1998; Bininda-Emonds and Sanderson, 2001; Day et al., 2008).

3.2.2. Parametric methods

Various parametric methods have been developed in the context of the coalescent. Carstens and Knowles (2007) suggest a coalescent approach to estimate phylogenies from the gene trees estimated from multilocus sequences. The likelihood scores of all possible phylogenies are calculated with the probability distribution of the gene tree topology given the species phylogeny (Degnan and Salter, 2005) and the best fit species tree is chosen by a likelihood ratio test with a correction for multiple comparisons (Anisimova and Gascuel, 2006).

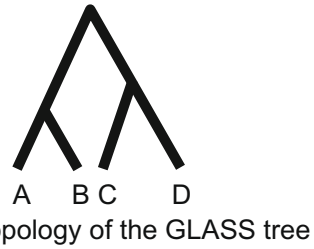
Coalescent theory assumes that gene coalescence times always predate species divergence times. This observation motivates the Global LAteSt Split (GLASS) method (Mossel and Roch, 2007) (also called the Maximum Tree (Liu et al., 2008; Liu et al., 2009a) which clusters species using minimum coalescences. This GLASS algorithm is based on an extension of Takahata's (1989) principle of minimum coalescence times to the collapsed gene tree concept of Rosenberg (2002). Given a collection of gene trees (Fig. 3a), GLASS first calculates the minimum gene coalescence times for all pairs of species across genes and then uses the minimum gene coalescence times to build an ultrametric tree (Fig. 3b). GLASS can be extended to use molecular distances instead of coalescence times to construct species phylogenies under the assumption that the rate of mutation is the same for all genes and all individuals in the same branch of the species phylogeny (Mossel and Roch, 2007). The principle of clustering species by minimum coalescence times is also implemented in the software Species Tree Estimation using Maximum Likelihood, or STEM (Kubatko et al., 2009). This method appears to perform well under a molecular clock and when rates among loci are equal.

Under the coalescence model, Liu et al. (2009b) proposed estimating species trees using average ranks of gene coalescence times (STAR). For the STAR method, the root has the highest rank and the



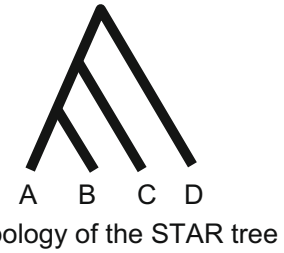
(b)

	A	B	C	D
A	--	0.2	0.6	0.8
B	0.2	--	0.6	0.8
C	0.6	0.6	--	0.3
D	0.8	0.8	0.3	--



(c)

	A	B	C	D
A	--	4.66	6.66	8
B	4.66	--	6.66	8
C	6.66	6.66	--	7.34
D	8	8	7.34	--



(d)

	A	B	C	D
A	--	0.86	1.86	2.2
B	0.86	--	1.86	2.2
C	1.86	1.86	--	1.4
D	2.2	2.2	1.4	--

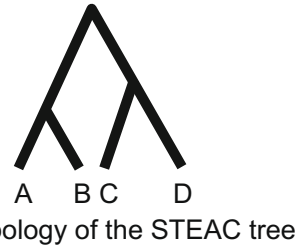


Fig. 3. The GLASS/Maximum tree, STAR, and STEAC methods of species tree inference. (a) A collection of gene trees: (((A:0.5, B:0.5):0.2, C:0.7):0.3, D:1.0); ((A:0.6, B:0.6):0.9, (C:0.3, D:0.3):1.2); (((A:0.2, B:0.2):0.4, C:0.6):0.2, D:0.8) branch lengths are in coalescent units. The numbers at the internodes in the gene trees are the ranks of the nodes. These gene trees assume a molecular clock but parametric methods often perform well in the absence of a clock (Liu et al., 2009b). (b) The matrix of minimum coalescence times across gene trees and the GLASS tree constructed from the minimum coalescence times ((A:0.2, B:0.2):0.4, (C:0.3, D:0.3):0.3). (c) The distance matrix of average ranks across gene trees and the STAR tree constructed from the distance matrix (((A:2.3, B:2.3):1, C:3.3):0.5, D:3.8). (d) The distance matrix of average coalescence times across gene trees and the STEAC tree constructed from the distance matrix ((A:0.43, B:0.43):0.58, (C:0.7, D:0.7):0.31).

rank decreases by one as it goes from the root to the leaves of the tree (Fig. 3a). The STAR method is implemented by building a distance tree, such as a Neighbor Joining (NJ) tree (Saitou and Nei, 1987) from the distance matrix in which the entries are twice the average ranks across gene trees (Fig. 3c). Liu et al. also proposed another method for species trees estimation using average coalescence times (STEAC). The species tree is estimated by a distance tree built from the distance matrix in which the entries are twice the average coalescence times (Fig. 3d) across gene trees. Whereas the STAR approach yields only species tree topologies, the STEAC methods yields consistent but upwardly biased branch lengths. In preliminary comparisons, STAR appears more robust than STEAC, while both methods are much faster than the Bayesian approach.

In an approach similar to consensus methods, Bayesian concordance factors (Ané et al., 2007; Baum, 2007) have been used to obtain information about the true underlying species relationships on a genome-wide scale. This method provides a measure of support for each potential species tree clade using a two-stage MCMC procedure in which posterior distributions of single-gene phylogenies

estimated in the first step are then used to form an estimate of the posterior distribution of the gene-to-tree maps in the second step. These gene-to-tree maps can then be summarized to provide information on the level of concordance for each clade across the multilocus gene trees while allowing for dependence between loci due to their shared species-level history. A notable feature of the method is that it does not require explicit specification of the underlying evolutionary process which generated the gene tree incongruence.

3.3. Comparison of methods for estimating species trees

3.3.1. Comparison between likelihood and summary statistic methods

Likelihood methods, including the maximum likelihood and Bayesian methods outlined in Section 3.1, make use of the full data to infer phylogenies. By contrast, the methods based on summary statistics estimate the phylogeny using only the summary statistic of the estimated gene trees. If the summary statistic is not sufficient, such methods use only partial information of the data to infer phylogenies and thus require more sequences (or genes) than the likelihood methods to achieve a certain level of confidence on the estimate of the species tree. However, it may not be possible to apply likelihood methods to large datasets such as seen more commonly in phylogenomic projects; such data sets often contain hundreds of genes and species trees are therefore challenging to estimate due to the intensive computation involved. By contrast, summary statistic methods can quickly infer phylogenies even for large-scale phylogenomic data (Liu et al., 2009b). Since phylogenomic data contains a huge amount of information regarding the phylogeny of species, the statistical efficiency of the methods for estimating phylogeny may not be the main concern; given the increasing size of phylogenomic data sets, computational ease may quickly become paramount.

3.3.2. Comparison among nonparametric methods

The consensus, reconciliation, and supertree methods estimate phylogenies using only the information contained in the topologies of the individual gene trees. Branch lengths of the gene trees estimated from the multilocus sequences are ignored by these nonparametric methods, and some methods ignore error in the gene trees. Degnan et al. (2008) studied the performance of several consensus methods in this setting, including 50% majority-rule consensus trees, R^* consensus trees, and greedy consensus trees. Bryant (2003) describes each of these consensus methods in detail, so here we give only a brief definition of each. Majority-rule consensus trees are formed by displaying all clades occurring with frequency 50% or more among the single gene tree estimates. R^* consensus trees are constructed by considering the three possible rooted phylogenetic relationships among each set of three taxa. Whenever a particular one of these three possible phylogenies of three taxa occurs in higher frequency in the gene trees than either of the other two, it is included in the R^* consensus tree. The greedy consensus tree is constructed by sequentially adding the most highly supported clades to the tree until a completely bifurcating tree results, with ties broken arbitrarily. Degnan et al. (2008) considered both asymptotic (in the number of genes) and finite-sample properties of each of these methods when the sole source of incongruence in the single-gene phylogenies is incomplete lineage sorting. They found that the majority-rule method was never inconsistent as an estimator of the species tree, but that it was unresolved sometimes, particularly in cases of anomalous or near-anomalous gene trees (Degnan and Rosenberg, 2006). R^* consensus trees were shown to be statistically consistent as the number of genes increases, but the convergence rate was found to be exceptionally slow, limiting their applicability for real data sets. Greedy consensus trees were fastest to converge, but often resulted in incorrect relationships among species. These results indicate that no single approach can outperform all other approaches in terms of both accuracy and speed. The situations that are difficult for consensus methods (for instance, estimating species trees in the anomaly zone) would also be difficult for other approaches such as parametric methods. However, there has not been a systematic comparison between nonparametric approaches and parametric approaches.

Although the issue of consistency has not been addressed for reconciliation methods, Steel and Rodrigo (2008) address consistency for supertree methods. Their model (exponential error model) is based on an error function such that the species tree is based on a weighted function of gene trees that have been assessed as to their deviation from an hypothesized underlying species tree. They have shown that the maximum likelihood approach for combining gene trees into a species tree is statistically consistent even in the anomaly zone, while commonly used supertree methods such as matrix representation with parsimony can be statistically inconsistent under the exponential error model.

3.3.3. Comparison among parametric methods

GLASS is statistically consistent under the assumptions in Section 2 (Liu et al., 2009a). As the number of genes increases, the probability that the GLASS tree is congruent with the true phylogeny converges to 1.0 at an exponential rate. Since the phylogeny is determined by the minimum coalescent times in the GLASS method, the systematic bias of the minimum coalescence times can result in the wrong estimate of the phylogeny, for example, if the assumptions in Section 2 are not satisfied (such as when sequences are affected by horizontal transfer or hybridization). Such processes will change the order of the minimum coalescence times and GLASS will produce an incorrect estimate of the phylogeny. Thus, the GLASS method is not robust to biological events such as horizontal transfer and hybridization that may have occurred during the evolutionary process generating the sequences in the data.

Under the coalescence model, the STAR and STEAC methods are statistically consistent (Liu et al., 2009b). In addition, both methods are fairly robust to a limited amount of horizontal transfer as well as deviations from a molecular clock because some small values of coalescence times due to horizontal transfer or rate variation in particular genes do not have major effects on the average ranks and average coalescence times when the number of genes is moderate or large.

4. Future directions

4.1. Extending the coalescence model

The methods for estimating phylogenies described in this review are based on the coalescence model that assumes no horizontal transfer and no gene flow among species. This model is probably appropriate for many clades where HGT is not common, such as higher eukaryotes, or when taxa are sampled such that gene flow is not a confounding variable (Liu et al., 2008). To make it applicable to a broader class of data that may have undergone horizontal transfer or gene flow, the coalescence model must be extended to accommodate these biological factors (Eckert and Carstens, 2008). There have been a variety of methods to detect the occurrence of HGT. The phylogeny-based approaches identify HGT by finding a phylogenetic network in which a minimum number of HGT events are required to reconcile gene trees and the species tree (Hallett and Lagergren, 2001; Nakhleh et al., 2005b; Beiko and Hamilton, 2006). Alternatively, the parsimony-based HGT detection approaches attempt to find a phylogenetic network that can minimize the parsimony length of the sequences evolved on

the network (Nakhleh et al., 2005a; Jin et al., 2006, 2007). Recently, Than et al. (2008) developed a new approach that can integrate parsimony-based approaches and phylogeny-based approaches to achieve highly efficient performance in terms of both accuracy and speed. In addition, several authors have recently attempted to consider both coalescence and hybridization as causes for observed discord in the histories of individual genes using various approaches (Buckley et al., 2006; Maureira-Butler et al., 2008; Meng and Kubatko, 2009).

The coalescent model described above can be extended in a variety of ways by adding more parameters and integrating more biological phenomena. For example, one area in major need of further research is estimating species trees in the absence of a molecular clock, something that was accomplished long ago for gene trees. Currently several species tree methods assume that the root of the species tree is known (for example, BEST (Liu, 2008) and STAR) and it would be useful to relax this assumption, as well as to devise ways of allowing the species tree to be non-ultrametric. Some of our own research is currently directed toward this end. Such extensions necessarily involve adding more parameters to the basic model. However, simply adding more parameters does not mean that the model is better. If the model contains too many parameters, it will eventually become inestimable because the data do not have sufficient information for estimating these parameters. Even the current Bayesian model (Liu et al., 2008) has quite a few parameters for moderate data sets. For example, in a species tree of 5 species for which 5 genes have been sampled, there are approximately 63 parameters, whereas for a typical Bayesian concatenated analysis of the same data, there are only 13 parameters under the Jukes–Cantor substitution model. Because of the multiple levels of analysis of both gene and species trees, species tree estimation necessarily involves more parameters, yet the most computationally advantageous models may be those that minimize the number of parameters to be estimated.

4.2. Checking model assumptions and goodness-of-fit for the coalescence model

The statistical properties of the methods described in this review are based on the assumption that the coalescence model is an accurate representation of the underlying population processes. For real problems, the actual process never exactly follows the expectation of any specific model. Some of the assumptions of the coalescence model may not fit the data. It is important to assess how well the model fits the data, realizing that a perfect fit is not expected. However, we do expect a good model to explain the data adequately. The fit of the multispecies coalescent model can be measured by the similarity between the gene trees estimated from data and those expected from the multispecies coalescent model. Further research is needed to find a sensitive measure to assess the similarity between two trees so that the deviation of the estimated gene trees from the expected ones can be used to detect significant departures from the multispecies coalescent model.

4.3. Guidance on data collection

The problem of "more loci, more alleles or more base pairs per locus" has been addressed by Felsenstein (2006) in the context of the estimation of population size θ. The same question may be asked by the researchers who are interested in the estimation of species trees. Preliminary work in this area was undertaken by Maddison and Knowles (2006), who examined the accuracy of the minimize deep coalescences and shallowest divergences methods at varying sampling efforts by measuring the proportion of correctly inferred clades in the species tree in a collection of simulated data sets. They found that the optimal allocation of sampling effort depended on properties of the true underlying species tree, with a direct effect due to the total depth of the species tree. At larger depths (on the order of $10N_e$), allocation of samples to more loci, rather than more individuals, was found to be most beneficial for a fixed total sampling effort, while at shallower depths (on the order of $1N_e$) allocation to the sampling of more individuals per species resulted in the greatest gains in accuracy. Whether these conclusions generalize to parametric or likelihood methods of species tree inference has yet to be examined, though we note that large scale simulations are often more difficult for these methods as they are generally more time-intensive. Care is also needed in selecting the measures of accuracy used to evaluate the estimated species tree. For example, some measures will capture differences in topology only, while others, such as the branch score distance (Kuhner and Felsenstein, 1994) also take branch lengths into account. Another useful approach to the problem would be to define a function to measure how much information is in the data and to identify the relationship between the number of loci, number of sequences per species (alleles), the length of loci in base pairs and the amount of information in the data. Since the variance of the estimate of the species tree will decrease as the information in the data increases, the variance of the estimator of the species tree can be used as such function to measure the amount of information contained in the data.

4.4. Combining different types of data

Molecular sequence data have been predominantly used in estimating phylogenies. Nevertheless, morphological, behavioral, and physiological traits also exhibit strong phylogenetic signal (Blomberg et al., 2003) since the evolutionary process of these traits are related to the phylogeny of species. Additionally, such traits may also experience the same kinds of genealogical discordance as molecular traits, due to the same kinds of processes. It is desirable to combine the information from different types of data to estimate phylogenies. The evolutionary process of species includes not only the changes of genetic material, but also the changes of morphological, behavioral, and physiological traits of species. Studies on combining different types of data can shed the light on the relationship of changes of genetic material and changes of morphological, behavioral, and physiological traits, and how these changes trigger speciation.

Acknowledgments

We thank Cecile Ané and David Baum for helpful discussion and Noah Rosenberg and an anonymous reviewer for helpful comments on the manuscript. This research is supported by National Science Foundation Grant DEB 0743616 to Scott Edwards and Dennis Pearl.

References

Adachi, J., Waddell, P.J., Martin, W., Hasegawa, M., 2000. Plastid genome phylogeny and a model of amino acid substitution for proteins encoded by chloroplast DNA. J. Mol. Evol. 50, 348–358.
Ané, C., Larget, B., Baum, D.A., Smith, S.D., Rokas, A., 2007. Bayesian estimation of concordance among gene trees. Mol. Biol. Evol. 24, 412–426.
Anisimova, M., Gascuel, O., 2006. Approximate likelihood-ratio test for branches: A fast, accurate, and powerful alternative. Syst. Biol. 55, 539–552.
Arvestad, L., Berglund, A.-C., Lagergren, J., Sennblad, B., 2003. Bayesian gene/species tree reconciliation and orthology analysis using mcmc. Bioinformatics 19, 7–15.
Avise, J.C., 2000. Phylogeography: the history and formation of species. Harvard University Press.
Baum, B.R., 1992. Combining trees as a way of combining data sets for phylogenetic inference, and the desirability of combining gene trees. Taxon 41, 3–10.
Baum, D.A., 2007. Concordance trees, concordance factors, and the exploration of reticulate genealogy. Taxon. 56, 417–426.
Beiko, R.G., Hamilton, N., 2006. Phylogenetic identification of lateral genetic transfer events. BMC Evol. Biol. 6, 15.

Berglund-Sonnhammer, A.C., Steffansson, P., Betts, M.J., Liberles, D.A., 2006. Optimal gene trees from sequences and species trees using a soft interpretation of parsimony. J. Mol. Evol. 63, 240–250.
Bininda-Emonds, O.R.P., Bryant, H.N., 1998. Properties of matrix representation with parsimony analyses. Syst. Biol. 47, 497–508.
Bininda-Emonds, O.R.P., Sanderson, M.J., 2001. Assessment of the accuracy of matrix representation with parsimony analysis supertree construction. Syst. Biol. 50, 565–579.
Blomberg, S.P., Garland, T., Ives, A.R., 2003. Testing for phylogenetic signal in comparative data: Behavioral traits are more labile. Evolution 57, 717–745.
Bonizzoni, P., Della Vedova, G., Dondi, R., 2003. Reconciling gene trees to a species tree. Algorithms and complexity. Springer, Berlin. pp. 120–131.
Brower, A.V.Z., DeSalle, R., Vogler, A., 1996. Gene trees, species trees, and systematics: a cladistic perspective. Annu. Rev. Ecol. Evol. S 27, 423–450.
Bryant, D., 2003. A classification of consensus methods for phylogenies. In: Janowitz, M., Lapointe, F.-J., McMorris, F.R., Mirkin, B., Roberts, F.S. (Eds.), BioConsensus, Center for Discrete Mathematics and Theoretical Computer Science. American Mathematical Society, Providence, RI, pp. 163–184.
Buckley, T.R., Cordeiro, M., Marshall, D.C., Simon, C., 2006. Differentiating between hypotheses of lineage sorting and introgression in New Zealand alpine cicadas (maoricicada dugdale). Syst. Biol. 55, 411–425.
Cao, Y., Janke, A., Waddell, P.J., Westerman, M., Takenaka, O., Murata, S., Okada, N., Paabo, S., Hasegawa, M., 1998. Conflict among individual mitochondrial proteins in resolving the phylogeny of eutherian orders. J. Mol. Evol. 47, 307–322.
Carstens, B.C., Knowles, L.L., 2007. Estimating species phylogeny from gene-tree probabilities despite incomplete lineage sorting: an example from melanoplus grasshoppers. Syst. Biol. 56, 400–411.
Casella, G., Berger, R.L., 2002. Statistical Inference. Duxbury Press.
Cotton, J.A., Wilkinson, M., 2007. Majority-rule supertrees. Syst. Biol. 56, 445–452.
Day, W.H.E., McMorris, F.R., Wilkinson, M., 2008. Explosions and hot spots in supertree methods. J. Theor. Biol. 253, 345–348.
Degnan, J.H., DeGiorgio, M., Bryant, D., Rosenberg, N.A., 2008. Properties of consensus methods for inferring species trees from gene trees. arxiv:/arXiv:0802.2355v1 [q-bio.PE].
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.
Doyle, J.J., 1992. Gene trees and species trees – molecular systematics as one-character taxonomy. Syst. Bot. 17, 144–163.
Eckert, A.J., Carstens, B.C., 2008. Does gene flow destroy phylogenetic signal? The performance of three methods for estimating species phylogenies in the presence of gene flow. Mol. Phylogenet. Evol. 49, 832–842.
Edwards, S.V., 2009. Is a new and general theory of molecular systematics emerging? Evolution 63, 1–19.
Eulenstein, O., Mirkin, B., Vingron, M., 1998. Duplication-based measures of difference between gene and species trees. J. Comput. Biol. 5, 135–148.
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., 1988. Phylogenies from molecular sequences: Inference and reliability. Annu. Rev. Genet. 22, 521–565.
Felsenstein, J., 2004. Inferring Phylogenies. Sinauer Associates, Sunderland, Massachusetts.
Felsenstein, J., 2006. Accuracy of coalescent likelihood estimates: Do we need more sites, more sequences, or more loci? Mol. Biol. Evol. 23, 691–700.
Fisher, R.A., 1922. On the mathematical foundations of theoretical statistics. Phil. Trans. R. Soc. Lond. A 222, 309–368.
Gadagkar, S.R., Rosenberg, M.S., Kumar, S., 2005. Inferring species phylogenies from multiple genes: Concatenated sequence tree versus consensus gene tree. J. Exp. Zool. B 304B, 64–74.
Gorbunov, K.Y., Lyubetsky, V.A., 2005. Identification of ancestral genes introducing incongruence between gene and species trees. Mol. Biol. 39, 847–858.
Hallett, M.T., Lagergren, J., 2001. Efficient algorithms for lateral gene transfer problems. Proc. 5th Ann. Int'l Conf. Comput. Mol. Biol., New York 14, 9–156.
Hasegawa, M., Kishino, H., Yano, T., 1985. Dating of the human-ape splitting by a molecular clock of mitochondrial DNA. J. Mol. Evol. 22, 160–174.
Hastings, W., 1970. Monte carlo sampling methods using markov chain and their applications. Biometrika 57, 97–109.
Hillis, D.M., Allard, M.W., Miyamoto, M.M., 1993. Analysis of DNA sequence data: phylogenetic inference. Method Enzymol. 224, 456–487.
Holland, B.R., Benthin, S., Lockhart, P.J., Moulton, V., Huber, K.T., 2008. Using supernetworks to distinguish hybridization from lineage-sorting. BMC Evol. Biol. 8, 202.
Hudson, R.R., 1992. Gene trees, species trees and the segregation of ancestral alleles. Genetics 131, 509–512.
Huelsenbeck, J.P., Bull, J.J., Cunningham, C.W., 1996. Combining data in phylogenetic analysis. Trends Ecol. Evol. 11, 152–158.
Huson, D.H., Klopper, T., Lockhart, P.J., Steel, M.A., 2005. Reconstruction of reticulate networks from gene trees. In: Research in Computational Molecular Biology, Proceedings. pp. 233–249
Jin, G., Nakhleh, L., Snir, S., Tuller, T., 2006. Efficient parsimony-based methods for phylogenetic network reconstruction. Bioinformatics 23, e123–e128. Proceedings of the European Conference on Computational Biology.
Jin, G., Nakhleh, L., Snir, S., Tuller, T., 2007. Inferring phylogenetic networks by the maximum parsimony criterion: a case study. Mol. Biol. Evol. 24, 324–337.
Jukes, T.H., Cantor, C.R., 1969. Evolution of Protein Molecules. Mammalian Protein Metabolism. Academic Press, New York. pp. 21–123.
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.
Kubatko, L., Carstens, B.C., Knowles, L.L., 2009. STEM: Species tree estimation using maximum likelihood for gene trees under coalescence. Bioinformatics 25, 971–973.
Kuhner, M.K., Felsenstein, J., 1994. A simulation comparison of phylogeny algorithms under equal and unequal evolutionary rates. Mol. Biol. Evol. 11, 459–468.
Linz, S., Radtke, A., von Haeseler, A., 2007. A likelihood framework to measure horizontal gene transfer. Mol. Biol. Evol. 24, 1312–1319.
Liu, L., 2008. BEST: Bayesian estimation of species trees under the coalescent model. Bioinformatics 24, 2542–2543.
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., 2009a. Maximum tree: A consistent estimator of the species tree. J. Math. Biol. 10.1007/s00285-009-0260-0.
Liu, L., Yu, L., Pearl, D.K., Edwards, S.V., 2009b. Estimating species phylogenies using coalescence times among sequences. Syst. Biol., doi:10.1093/sysbio/syp031.
Ma, B., Li, M., Zhang, L.X., 2000. From gene trees to species trees. SIAM J. Comput. 30, 729–752.
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.
Margush, T., McMorris, F.R., 1981. Consensus *n*-trees B. Math. Biol. 43, 239–244.
Maureira-Butler, I.J., Pfeil, B.E., Muangprom, A., Osborn, T.C., Doyle, J.J., 2008. The reticulate history of medicago (fabaceae). Syst. Biol. 57, 466–482.
Meng, C., Kubatko, L.S., 2009. Detecting hybrid speciation in the presence of incomplete lineage sorting using gene tree incongruence. A model. Theor. Popul. Biol. 75, 35–45.
Mossel, E., Roch, S., 2007. Incomplete lineage sorting: consistent phylogeny estimation from multiple loci. Available from: http://arxiv.org/abs/0710.0262.
Nakhleh, L., Jin, G., Zhao, F., Mellor-Crummey, J., 2005a. Reconstructing phylogenetic networks using maximum parsimony. In: Proceedings of the 2005 IEEE Computational Systems Bioinformatics Conference. pp. 93–102.
Nakhleh, L., Ruths, D., Wang, L.S., 2005b. A fast and accurate heuristic for reconstructing horizontal gene transfer. In: Wang, L. (Ed.), Cocoon 2005. Lncs. Springer, Heidelberg, pp. 84–93.
Nichols, R., 2001. Gene trees and species trees are not the same. Trends Ecol. Evol. 16, 358–364.
Nishihara, H., Okada, N., Hasegawa, M., 2007. Rooting the eutherian tree: the power and pitfalls of phylogenomics. Genome Biol. 8, R199.
Page, R.D.M., 1998. Genetree: comparing gene and species phylogenies using reconciled trees. Bioinformatics 14, 819–820.
Page, R.D.M., 2000. Extracting species trees from complex gene trees: Reconciled trees and vertebrate phylogeny. Mol. Phylogenet. Evol. 14, 89–106.
Page, R.D.M., Charleston, M.A., 1997. From gene to organismal phylogeny: reconciled trees and the gene tree species tree problem. Mol. Phylogenet. Evol. 7, 231–240.
Pamilo, P., Nei, M., 1988. Relationships between gene trees and species trees. Mol. Biol. Evol. 5, 568–583.
Pollard, D.A., Iyer, V.N., Moses, A.M., Eisen, M.B., 2006. Widespread discordance of gene trees with species tree in drosophila: evidence for incomplete lineage sorting. Plos Genet. 2, 1634–1647.
Powell, J.R., 1991. Monophyly/paraphyly/polyphyly and gene/species trees - an example from drosophila. Mol. Biol. Evol. 8, 892–896.
Rannala, B., Yang, Z., 2008. Phylogenetic inference using whole genomes. Annu. Rev. Genom. Hum. Genet. 9, 217–231.
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.
Rosenberg, N.A., 2002. The probability of topological concordance of gene trees and species trees. Theor. Popul. Biol. 61, 225–247.
Saitou, N., Nei, M., 1987. The neighbor-joining method – a new method for reconstructing phylogenetic trees. Mol. Biol. Evol. 4, 406–425.
Sanderson, M.J., McMahon, M.M., 2007. Inferring angiosperm phylogeny from est data with widespread gene duplication. BMC Evol. Biol. 7 (Suppl 1), S3.
Seo, T.K., Kishino, H., Thorne, J.L., 2005. Incorporating gene-specific variation when inferring and evaluating optimal evolutionary tree topologies from multilocus sequence data. Proc. Natl. Acad. Sci. USA 102, 4436–4441.
Slowinski, J.B., Knight, A., Rooney, A.P., 1997. Inferring species trees from gene trees: a phylogenetic analysis of the elapidae (serpentes) based on the amino acid sequences of venom proteins. Mol. Phylogenet. Evol. 8, 349–362.
Steel, M., Rodrigo, A., 2008. Maximum likelihood supertrees. Syst. Biol. 57, 243–250.
Stege, U., 1999. Gene trees and species trees: The gene-duplication problem is fixed-parameter tractable. Algorithms and data structures. pp. 288–293

Sullivan, J., 2005. Maximum-likelihood methods for phylogeny estimation. Methods Enzymol. 395, 757–779.
Swofford, D.L., Olsen, G.J., Waddell, P.J., Hillis, D.M., 1996. Phylogenetic Inference. In: D.M. Hillis, C. Moritz, B.K. Mable, (Eds.), Molecular Systematics. Sinauer, Sunderland, MA.
Takahata, N., 1989. Gene genealogy in three related populations: Consistency probability between gene and population trees. Genetics 122, 957–966.
Than, C., Jin, G., Nakhleh, L., 2008. Integrating sequence and topology for efficient and accurate detection of horizontal gene transfer. In: Proceedings of the Sixth RECOMB Comparative Genomics Satellite Workshop. Lecture Notes in Bioinformatics. pp. 113–127.
V'Yugin, V.V., Gelfand, M.S., Lyubetsky, V.A., 2002. Tree reconciliation: reconstruction of species phylogeny by phylogenetic gene trees. Mol. Biol. 36, 650–658.
Wakeley, J., 2008. Coalescent Theory: An Introduction. Roberts & Company Publishers.
Whelan, S., Goldman, N., 2001. A general empirical model of protein evolution derived from multiple protein families using a maximum likelihood approach. Mol. Biol. Evol. 18, 691–699.
Wilkinson, M., Cotton, J.A., Creevey, C., Eulenstein, O., Harris, S.R., Lapointe, F.J., Levasseur, C., McInerney, J.O., Pisani, D., Thorley, J.L., 2005. The shape of supertrees to come: tree shape related properties of 14 supertree methods. Syst. Biol. 54, 419–431.