On the Robustness to Gene Tree Estimation Error (or lack thereof) of Coalescent-Based Species Tree Methods

SEBASTIEN ROCH1,* AND TANDY WARNOW2

1Department of Mathematics, University of Wisconsin at Madison, 480 Lincoln Dr., Madison, Wisconsin, 53706, USA and 2Departments of Bioengineering and Computer Science, University of Illinois at Urbana-Champaign, Urbana, IL, 61801, USA

*Correspondence to be sent to: Department of Mathematics, The University of Wisconsin at Madison, USA; E-mail: roch@math.wisc.edu

Received 11 December 2014; reviews returned 16 February 2015; accepted 20 March 2015

Associate Editor: Laura Kubatko

Abstract.—The estimation of species trees using multiple loci has become increasingly common. Because different loci can have different phylogenetic histories (reflected in different gene tree topologies) for multiple biological causes, new approaches to species tree estimation have been developed that take gene tree heterogeneity into account. Among these multiple causes, incomplete lineage sorting (ILS), modeled by the multi-species coalescent, is potentially the most common cause of gene tree heterogeneity; and much of the focus of the recent literature has been on how to estimate species trees in the presence of ILS. Despite progress in developing statistically consistent techniques for estimating species trees when gene trees can differ due to ILS, there is substantial controversy in the systematics community as to whether to use the new coalescent-based methods or the traditional concatenation methods. One of the key issues that has been raised is understanding the impact of gene tree estimation error on coalescent-based methods that operate by combining gene trees. Here we explore the mathematical guarantees of coalescent-based methods when analyzing estimated rather than true gene trees. Our results provide some insight into the differences between promise of coalescent-based methods in theory and their performance in practice. [coalescent-based methods; gene tree estimation error; incomplete lineage sorting; multi-species coalescent; species tree reconstruction; statistical consistency]

With the increasing ease in generating sequence data (and even whole genome assemblies), phylogenetic analyses are being attempted on large numbers of multiple loci. Although some of these phylogenetic analyses seem uncontroversial, others present substantial challenges, often as a result of conflict between different gene trees. While several biological processes can lead to this gene tree conflict (e.g., gene duplication and loss, incomplete lineage sorting [ILS], and horizontal gene transfer), ILS (modeled by the multi-species coalescent) is considered to be one of the major causes for gene tree conflict (Edwards 2009).

Because the standard approach of using maximum likelihood on concatenated alignments is statistically inconsistent under the multi-species coalescent model (Roch and Steel 2014) and can result in highly supported but incorrect trees in simulation studies (Kubatko and Degnan 2007), alternative species tree estimation methods that explicitly take gene tree discordance into account have been developed. Some of the most promising of these coalescent-based methods co-estimate gene trees and species trees from input sequence alignments using Bayesian MCMC techniques (e.g., *BEAST; Heled and Drummond 2010 and BEST; Liu 2008), but these co-estimation methods are rarely attempted on datasets containing 100 or more loci or more than 30 or so species due to their computational requirements (Bayzid and Warnow 2013; McCormack et al. 2013; Zimmermann et al. 2014). Instead, most coalescent-based analyses of biological datasets are performed using summary methods that estimate species trees by combining gene trees. Many summary methods are statistically consistent, and so will return the true species tree given a large enough number of gene trees for loci randomly sampled from throughout the genome, and some of these methods (e.g., MP-EST [Maximum Pseudo-Likelihood for Estimating Species Trees]; Liu and Edwards 2010b) have become very popular.

Despite the progress in developing summary methods with statistical guarantees under the multi-species coalescent model, there is great controversy over whether these coalescent-based species tree estimation methods should be used, or whether the standard approach of concatenation should be used even though it is not statistically consistent. Perhaps the most compelling reason for preferring coalescent-based methods over concatenation is that they are provably statistically consistent, whereas concatenation is not. However, critiques of summary methods based on a combination of theoretical and empirical grounds have also been made (Gatesy and Springer 2013, 2014). One of the major concerns in these studies is the use of sequences taken from large genomic regions (whether contiguous or not) in coalescent-based analyses, since these regions have a high probability of involving some recombination and recombination violates one of the main assumptions of the multi-species coalescent model. (Gatesy and Springer 2014) surveyed the literature and found only six published studies in which concatenation and coalescent-based methods were compared on datasets generated under
the multi-species coalescent model (Kubatko et al. 2009; Liu et al. 2009a; Liu and Edwards 2010b; Leaché and Rannala 2011; Bayzid and Warnow 2013; Patel et al. 2013). Based on these studies, (Gatesy and Springer 2014) concluded that the simulation model conditions in which concatenation was less accurate were for small numbers of species, or used gene trees estimated on long sequences. They noted that limiting analyses to short enough regions to avoid recombination would produce very short sequence alignments and increase gene tree estimation error (GTEE), and that various summary methods had impaired accuracy when GTEE was high. Finally, (Gatesy and Springer 2014) also reanalyzed a 183-locus, 24-species mammalian dataset previously studied in (McCormack et al. 2012) that used relatively short sequences flanking ultra-conserved elements (UCEs), which were short enough to potentially avoid recombination within loci. Their reanalysis examined whether two summary methods, STAR (Liu et al. 2009b) and MP-EST (Liu and Edwards 2010b), were able to recover 10 well-established groups within placental mammals, and showed that coalescent-based analyses failed to recover some groups (e.g., Glires), although concatenation succeeded in recovering them with high support.

Thus, based on both theoretical grounds (violations of model assumptions) and empirical grounds (i.e., performance on both biological and simulated data), (Gatesy and Springer 2014) concluded that concatenation was preferable to coalescent-based summary methods, and called on systematists to avoid the use of long gene sequences within coalescent-based analyses, due to the increased probability of recombination within loci.

Three recent studies (Mirarab et al. 2014a,b,c) not discussed in (Gatesy and Springer 2014) not only confirm many of the trends observed by (Gatesy and Springer 2014), but also show some new observations, and so are worth discussing. Consistent with some trends observed in (Gatesy and Springer 2014), (Mirarab et al. 2014c) showed that GTEE resulting from insufficient phylogenetic signal increases the error in species trees estimated using MP-EST, and that under these conditions concatenation can be more accurate than MP-EST. However, (Mirarab et al. 2014c) also presented a technique (called “statistical binning”) for improving gene tree estimation, and showed that using statistical binning with MP-EST made it at least as accurate as concatenation (and often more so). Finally, (Mirarab et al. 2014b) evaluated the relative performance of MP-EST when using best maximum likelihood gene trees or multi-locus bootstrapping gene trees, and showed that when the number of genes is sufficiently large, using MP-EST with multi-locus bootstrapping (which is how most analyses of MP-EST have been performed) produced less accurate species trees than using best maximum likelihood trees, and conversely when the number of gene trees is sufficiently small using MP-EST with multi-locus bootstrapping gene trees produces better accuracy than using best maximum likelihood trees. Finally, (Mirarab et al. 2014a) introduced a new summary method, Accurate Species Tree Algorithm (ASTRAL), and showed that ASTRAL and MP-EST were both more accurate than concatenation under a wide range of model conditions. Thus, new studies confirm some conclusions made by (Gatesy and Springer 2014), but in general more nuanced: They show that the relative performance of summary methods and concatenation may depend not only on the model condition and sequence lengths (the main focus of Gatesy and Springer 2014), but also on the specific summary method and how it is run (i.e., with best maximum likelihood gene trees or multi-locus bootstrapping gene trees).

Furthermore, they present new coalescent-based species tree estimation techniques (MP-EST with statistical binning and ASTRAL) that have greater robustness to GTEE, and seem to be reliably at least as accurate as concatenation under most tested conditions.

However, we agree with the point made by (Gatesy and Springer 2014): If gene trees are based on long sequences then recombination within loci is likely to occur, whereas if they are based on short sequences then GTEE is likely to occur. Hence, one way or another, summary methods will likely be applied to gene trees that violate at least one of these two model assumptions (that gene trees are estimated correctly, and that each gene evolves without recombination). Therefore, it makes sense to ask whether current summary methods have any provable statistical properties in the presence of recombination within loci, or on data sets with GTEE. Here, we focus on the second of these questions: Whether summary methods are statistically consistent in the presence of GTEE.

A careful examination of the proofs of statistical consistency for summary methods reveals that generally summary methods have only been proven to converge to the true species tree given perfect gene trees without any estimation error. One notable exception is the work of (Steel and Rodrigo 2008) who prove the consistency of a maximum likelihood supertree approach under a special model of GTEE with an exponential form (under a strict molecular clock). With the exception of that work, all proofs of statistical consistency under the multi-species coalescent model for summary methods have the following form: For a given model species tree and a given species tree reconstruction error probability $\epsilon > 0$, there is a number $K$ so that with probability at least $1-\epsilon$, the method will return the true species tree when given at least $K$ true gene trees sampled randomly from the distribution. Thus, the proofs inherently require true gene trees, and do not provide a guarantee of robustness to GTEE. On the contrary, that does not mean that the methods are not able to reconstruct the true tree with high probability when gene trees have estimation error—it only means that we do not yet have any proofs of mathematical guarantees for these coalescent-based species tree estimation under conditions with GTEE. Furthermore, for many model conditions, simulation studies have shown that highly accurate species trees can be constructed using summary methods even in the presence of GTEE (Larget et al. 2010; Liu and...
Edwards 2010b; Bayzid and Warnow 2013; Mirarab et al. 2014a,b). However, because some coalescent-based summary methods sometimes produce less accurate estimates of species trees than concatenation (Bayzid and Warnow 2013; Patel et al. 2013; Mirarab et al. 2014c), seemingly as a result of GTEE, it makes sense to ask whether there are any provable guarantees that can be established for summary methods in the presence of GTEE.

We approach this question by considering the theoretical guarantees of summary methods when the number of genes can grow, but the sequence length for every gene is bounded by a constant. This approach is biologically motivated if the individual loci are restricted to genomic regions that are short enough to be likely to be recombination-free. Since the length of these recombination-free regions may be quite short, especially for deep evolutionary scenarios or large data sets, and may depend on the organism (Hohlohl et al. 2011; Gatesy and Springer 2014), this is a matter of substantial concern to some investigators. In particular, we ask whether we can develop techniques for combining gene trees that will converge to the true species tree when the gene trees have some estimation error. To make this study biologically meaningful, we model GTEE as a consequence of limited sequence length.

The theoretical framework and results are presented below. When gene sequence evolution occurs under a strict molecular clock (so that the expected number of changes per site is proportional to time), we present a new species tree estimation method that is statistically consistent even when each gene is represented by a single site (Corollary 1). For the case where the strict molecular clock assumption is violated, we present a species tree estimation method that is statistically consistent when all the gene sequence alignments are exponentially long in the model species tree (Theorem 3). Both results are proved under the maximum evolutionary distance in the model species tree (Corollary 1). For the case where the strict molecular clock is not assumed, then statistical consistency in the presence of GTEE is possible using very similar methods. For example, we will show that the following approach is statistically consistent:

- Step 1: For each gene and for every four species, compute an estimated quartet tree with the Four Point Method (e.g., as in Erdős et al. 1999a,b) using appropriate distance estimates.
- Step 2: For every four species, compute the most frequent quartet tree from among the three possibilities using the quartet trees computed in Step 1 (breaking ties randomly).
- Step 3: If the quartet trees are compatible, return the unique unrooted tree that is consistent with them; otherwise return Fail.

Another method described below is also statistically consistent under these assumptions, and has the desirable property that it remains statistically consistent at all sequence lengths. In fact, the main requirement is that all triplet gene trees are estimated using a method that is biased, however slightly, toward the true rooted gene tree on each triplet of species. What this means is that if the true gene tree on species set \( \{a, b, c\} \) is \( ab|c \) (so \( a \) and \( b \) are siblings in the gene tree), then the probability that the gene tree estimation method returns \( ab|c \) on sequences of finite length is strictly greater than the probability of returning one of the other two trees. Note that this bias requirement can be achieved with \( Pr(ab|c) = 1/3 + 2\varepsilon \) and \( Pr(ac|b) = Pr(bc|a) = 1/3 - \varepsilon \), and so a very small bias suffices for this theoretical result.

For the case where the strict molecular clock is not assumed, then statistical consistency in the presence of GTEE is possible using very similar methods. For example, we will show that the following approach is statistically consistent:

- Step 1: For each gene and for every three species, estimate the rooted gene tree by making the two species with the smallest Hamming distance siblings; these are the rooted triplet trees (if there are ties in the Hamming distances, then randomly break the tie).
- Step 2: For every three species, select the rooted gene tree that appears the most frequently among the three possible rooted gene trees for that set of three species, from the rooted triplet trees computed in Step 1. This is the “dominant” triplet tree for the three species. (As before, if there is a tie for the most frequent rooted three-leaf gene tree for a given three species, break the tie randomly.)
- Step 3: Determine if the set of dominant triplet trees is compatible using (Aho et al. 1981). If so, return the unique tree that agrees with the set. Otherwise, return Fail.

Under the Jukes–Cantor model of sequence evolution, this method is statistically consistent, so that as the number of genes increases the probability of returning the true species tree goes to 1. However, for the proof to work, the probability of computing the true quartet gene tree must be very close to 1 for nearly all genes. This is in contrast to the rooted case, where it sufficed to use any gene tree estimation method that was biased toward the true gene tree. Therefore, while this method is guaranteed to return the true species tree with high probability given a large enough number of genes, nearly all genes must come with long enough sequences, and the length that is needed for the proof to work increases as the length of the shortest branch in the species...
The difference between the case with the strict molecular clock and the case where there is no strict molecular clock is substantial, especially since the number of sites needed for the proof can be very large in the presence of substantial ILS (which results from very short internal branches in the species tree) or deep evolutionary histories (which can result in species trees and gene trees with high evolutionary diameters). These differences are biologically significant, since gene trees may not generally obey the strict molecular clock.

Furthermore, although we provide proofs of accuracy with high probability for some methods, no positive results have been established for any of the standard coalescent-based methods (such as MP-EST (Liu and Edwards 2010), the population tree from BUCKy (Larget et al. 2010), and ASTRAL (Mirarab et al. 2014a)) that operate by combining gene trees estimated on the full set of species. The difficulty in providing proofs of accuracy for species tree estimation methods that do not estimate quartet gene trees or triplet gene trees independently, but rather compute a gene tree on all the taxa and then examine induced triplet trees or quartet trees, is that modeling GTEE that arises in this process is much more challenging. Thus, although it is possible that these other methods are also statistically consistent in the presence of GTEE, there is currently no proof that the true species tree will be returned with high probability by ASTRAL, MP-EST, etc., given a large enough number of gene trees, when they may all have some small probability of error.

Overall, this study shows that the guarantees that have been previously established for coalescent-based methods that operate by combining gene trees are quite limited. While improved performance for some coalescent-based summary methods relative to concatenation has been established in some empirical and simulation studies, the vulnerability to GTEE has also been demonstrated. Further research is needed to establish the theoretical guarantees of existing summary methods with respect to vulnerability or robustness to GTEE, and to develop new coalescent-based summary methods that are provably robust to GTEE.

Recommendations

Several papers have presented arguments either for or against the use of coalescent-based summary methods (Gatesy and Springer 2014; Song et al. 2012) on the basis of both theory and empirical performance on data. However, the evidence supporting these arguments is incomplete, and relative performance on finite data clearly depends on the model conditions. Hence, we do not make any general recommendation in favor of one type of method or another.

In fact, the comparison is complicated even with respect to statistical consistency guarantees. If gene trees are all correctly estimated, then summary methods have the advantage as they are statistically consistent under the multi-species coalescent, whereas concatenation using an unpartitioned maximum likelihood analysis is not only inconsistent, but also can be positively misleading (Roch and Steel 2014). This study also showed that some summary methods can be statistically consistent in the presence of GTEE, but that guarantee required the strict molecular clock. When the strict molecular clock is violated and there is GTEE, then we have a very weak positive result—we introduced methods that are statistically consistent, but proved this only when nearly perfect gene trees are provided. (Given the sequence lengths needed to provide this guarantee and the increased probability of recombination within these sequences, the result is perhaps not of great practical relevance.) It is possible that some of the standard coalescent methods also are statistically consistent in the presence of GTEE, but our results do not provide these proofs. Thus, in terms of statistical consistency, we know that concatenation can be positively misleading in the presence of ILS, and we know that coalescent-based methods are statistically consistent when given true gene trees, but we have generally only a little bit of information about statistical consistency in the presence of GTEE. The choice between coalescent-based summary methods and concatenation on the basis of statistical consistency is, therefore, not at all straightforward.

With respect to accuracy in simulation studies, results are also mixed. As noted by (Gatesy and Springer 2014), many simulation studies have suggested that summary methods are less accurate than concatenation when given estimated gene trees. However, recent studies (Bayzid and Warnow 2013; Mirarab et al. 2014a,b) have shown conditions in which some coalescent-based methods can have better performance than concatenation on estimated gene trees. For example, Figure 3 in (Mirarab et al. 2014b) shows that the average GTEE (AGE), which is controlled by the gene sequence alignment length, impacted the relative performance of concatenation and MP-EST. Specifically, concatenation was generally less accurate than MP-EST given gene sequences of length 1000, very close in accuracy given sequences of length 500, and more accurate given gene sequences of length 250. Thus, gene sequence alignment length impacts the relative performance of concatenation and MP-EST, but MP-EST can be at least as accurate than concatenation even on relatively short sequences.

Another factor clearly impacting the relative performance of concatenation and summary methods is the level of ILS. The simulation study in (Mirarab et al. 2014a) based on 200 mammalian 500 bp gene sequence alignments of varying levels of ILS showed concatenation was more accurate than ASTRAL and MP-EST for very low levels of ILS, but became less accurate than both methods for large enough levels of ILS (Figure 2C in Mirarab et al. 2014a). This may be expected, since as the amount of ILS decreases the gene tree incongruence also decreases, reducing the degree to which the model assumptions for the maximum
likelihood concatenated analysis are violated. Because this particular comparison was based on only 200 genes, it does not imply that concatenation would still be more accurate than these coalescent-based methods given a larger number of genes. However, (Mirarab et al. 2014a) suggests the possibility that for relatively low levels of ILS, concatenation may be as accurate as the current summary methods, for some data set conditions (perhaps relatively short sequences, with limited numbers of genes). There is also the question of the impact of data set size, since nearly all evaluations of coalescent-based species tree estimation methods have been limited to data sets with at most 50 species, and relative performance of methods can change as the number of species increases.

Thus, although several simulation studies have shown conditions in which concatenation is more accurate than leading summary methods, others have shown conditions in which the reverse is true. Clearly, the relative performance of concatenation using maximum likelihood and current summary methods (in particular MP-EST and ASTRAL) depends on the number of genes, the amount of ILS, the amount of GTEE, and possibly also other factors such as the number of species. The theoretical results established here also suggest the possibility that the relative performance may depend on the presence of a molecular clock and possibly the gene sequence evolution model. These observations also indicate that the choice between concatenation and a coalescent-based summary method will need to depend on the particular features of the data set, and in particular the evidence for low or high amounts of ILS or GTEE. Unfortunately, without knowing the true species tree and gene trees, evaluating the amount of ILS in the data or the amount of GTEE is quite difficult. For example, discordance between estimated gene trees may be due to GTEE or true gene tree discordance, and determining which is the case may be important when choosing between species tree estimation protocols.

Thus, although we agree with many of the concerns raised by (Gatesy and Springer 2014) regarding the use of current summary methods, we do not agree with the conclusion that concatenation should be preferred over all summary methods. Similarly, we do not agree with the assertion that others have made that all coalescent-based methods should be preferred over concatenation on the basis of statistical guarantees. Instead, we find that the choice between the two types of species tree estimation protocols cannot be justified on purely theoretical grounds (at least not based on the current research), and conclude that the assumed theoretical advantage of coalescent-based summary methods over concatenation may not be applicable in the presence of GTEE. However, we also find that the current body of simulation studies present a more nuanced comparison between concatenation and coalescent-based methods, and show conditions in which the best coalescent-based methods can be more accurate than concatenation even with relatively high amounts of GTEE due to short sequence alignments. Thus, from both an empirical and a theoretical perspective, the choice between the two modes of analysis is complicated, and neither dominates the other. (See also (Sanderson and Kim 2000) for a related discussion.)

Finally, and most importantly, the mixed performance of the current summary methods should not be taken as evidence that summary methods are inherently limited in accuracy. Instead, despite the awareness of gene tree incongruence as an important challenge for phylogenomic analysis (Maddison 1997; Degnan and Rosenberg 2009; Edwards 2009), methods for estimating species trees from multiple loci are still in their first generation. Nevertheless, there is a very active effort in developing new coalescent-based species tree estimation methods. For example, ASTRAL (Mirarab et al. 2014a), which was used to estimate a plant phylogeny (Wickett et al. 2014), and MP-EST with statistical binning, which was used to estimate an avian phylogeny (Jarvis et al. 2014), both showed reliable advantages over concatenation on simulation studies as well as on biological data sets (see also Bayzid et al. 2014b). Divide-and-conquer techniques have also shown the ability to speed up coalescent-based species tree methods and also make them more accurate (see (Zimmermann et al. 2014) for divide-and-conquer used with *BEAST, and (Bayzid et al. 2014a) for its use with MP-EST).

At the same time, the focus on summary methods both in this article and in phylogenomic studies (Song et al. 2012; Jarvis et al. 2014; Wickett et al. 2014) limits the discussion. Alternative approaches that estimate species trees without combining estimated gene trees have also been developed (DeGiorgio and Degnan 2010; Chifman and Kubatko 2014; Dasarathy et al. 2014). Although some of these methods have been shown to be statistically consistent, their performance on biological data is less well understood. Given the challenges in providing theoretical proofs of convergence to the true tree from very short sequences for summary methods (except under the unrealistic assumption of a strict molecular clock), it may be that alternative approaches that do not combine estimated gene trees may be more promising.

Thus, rather than being pessimistic about coalescent-based species tree estimation and, therefore, arguing in favor of concatenation (which can be statistically inconsistent), we are optimistic that the next generation of coalescent-based species tree methods will produce techniques that are provably statistically consistent, able to provide outstanding accuracy under realistic conditions, and also fast enough to be used on biological data sets with thousands of species and many thousands of loci. These methods may also be reliably more accurate than concatenation even under the conditions where current coalescent-based methods are less accurate than concatenation.
**Theoretical Results**

We present new summary methods that are guaranteed to converge to the true species tree, given a large enough number of genes with bounded sequence length. However, the positive results we obtain only apply to new methods we present in this article, and not to existing summary methods. Furthermore, the main positive result only holds under the strict molecular clock (in which case, even a very simple method is guaranteed to converge to the true species tree for the extreme case where each gene has only a single site). When the strict molecular clock cannot be assumed, then we can show that a simple summary method we propose will converge to the true species tree, but this proof requires that nearly perfect gene trees be available for nearly all genes, for the true species tree to be obtained with high probability.

These results indicate possible directions for developing improved coalescent-based species tree estimation methods, but also reflect the limitations in the theoretical guarantees for the existing summary methods. While performance in practice may be better than the theoretical guarantees suggest, these are cautionary results, and imply that for many biological methods, we require that nearly perfect gene trees be available for large enough number of genes with bounded sequence length. However, the positive results we obtain only necessarily correctly reconstructed, that is, we do not consider only the topology of this estimated gene tree. However, we do not assume that gene trees are necessarily correctly reconstructed, that is, we do not assume that the topologies of $G_i$ and $G_0$ coincide. Instead, we posit a model of GTEE. We refer to $[G]_{i\in I}$ as the estimated gene trees. We will consider two types of estimated gene trees—rooted gene trees, and unrooted gene trees, understood in context.

**True Gene Trees**

*Species tree.*—Consider $n$ taxa with a common evolutionary history described by a rooted species tree $S = (V, E)$ with leaf set $L$. Note that $|L| = n$. To each branch $e$ of $S$, we assign:

- $N_e$, the (haploid) population size on e (we assume that the population size remains constant along each branch);
- $t_e$, the number of generations on $e$;
- $\tau_e = 2\mu e$, the length of $e$ in coalescent time units;
- $\mu_e$, the lineage-specific mutation rate on $e$ in units of coalescent time.

We also let $B$ be the distribution of gene-specific mutation rates over $(0, +\infty)$.

*Gene trees.*—For each taxon $i$ and each locus $i \in I$, with $|I| = g$, we sample one allele. Each locus $i \in I$ has a genealogical history represented by a gene tree $G_i = (V_i, E_i)$ with leaf set $L_{i} = I$. Following (Rannala and Yang 2003), we assume that each gene tree $G_i$ is distributed according to the multi-species coalescent process: Looking backwards in time, on each branch of the species tree, the coalescence of any two alleles is exponentially distributed with rate 1 in coalescent time units, independently from all other pairs; whenever two branches merge in the species tree, we also merge the allele sets of the corresponding populations, that is, the coalescence proceeds on the union of both allele sets. More specifically, the density of the likelihood of a rooted gene tree under this model is

$$
\prod_{i=1}^{g} \prod_{e \in E_i} \exp \left( -\frac{O_e}{2} \left( \sigma_{e}^{O_e+1} - \sigma_{e}^{O_e} \right) \right) e^{-O_e/2} \prod_{n=1}^{m} \exp \left( -\frac{m}{2} \left( \sigma_{e}^{m} - \sigma_{e}^{m-1} \right) \right),
$$

where, for locus $i$ and branch $e$, $I_{e}^i$ is the number of lineages entering $e$, $O_{e}^i$ is the number of lineages exiting $e$, and $\sigma_{e}^{m}$ is the $m$-th coalescence time in $e$; for convenience, we let $\sigma_{e}^{0}$ and $\sigma_{e}^{m}$, $O_{e}^i$, be, respectively, the divergence times of $e$ and of its parent population.

Gene tree $G_i$ is also assigned a gene-specific mutation rate $\beta_i$ distributed according to $B$, which scales all branch lengths. We further assume that the $k$ loci $\{I_{e}^i\}_{e \in E_i}$ are independent. We will refer to the gene trees generated by this process as the true gene trees. This leads us to the issue of GTEE.

**Estimated Gene Trees and GTEE Model**

Assume first that a gene tree $\hat{G}_i = (\hat{V}_i, \hat{E}_i)$ is inferred for each locus $i$. For our purposes, it will be enough to consider only the topology of this estimated gene tree. However, we do not assume that gene trees are necessarily correctly reconstructed, that is, we do not assume that the topologies of $G_i$ and $\hat{G}_i$ coincide. Instead, we posit a model of GTEE. We refer to $[\hat{G}]_{i\in I}$ as the estimated gene trees. We will consider two types of estimated gene trees—rooted gene trees, and unrooted gene trees, understood in context.

GTEE model.—GTEE is impacted by the method used to estimate gene trees, the model gene tree parameters, and the length of the gene sequence alignment. In what follows, we will assume that the sequence length is fixed, so that the estimation error depends only on the model gene tree and the gene tree estimation method.

For each gene tree estimation method and sequence length, our model of GTEE is of the following form. Let $T_{\Theta}$ be the set of bifurcating trees with $n$ leaves and let $W_{\Theta}$ be the set of trees in $T_{\Theta}$ with positive edge weights. We denote by $\Delta(T_{\Theta})$ the set of probability distributions on $T_{\Theta}$. Then our GTEE model is a mapping $R: W_{\Theta} \rightarrow \Delta(T_{\Theta})$, that is, to each true gene tree $G$ in $W_{\Theta}$ we have a probability distribution $\mathcal{D}(G)$ over (topologies of) estimated gene trees in $T_{\Theta}$. We make the following natural assumption about our GTEE model:
• R is equivariant in the sense that permuting the leaf labels of the input true gene tree \( \mathcal{G} \) similarly permutes the leaf labels of the output distribution \( R(\mathcal{G}) \) of estimated gene trees (in other words, the output of the phylogenetic reconstruction method used to estimate gene trees is unaffected by the names of the species).

We will consider both unrooted and rooted cases, which we indicate with superscripts (u) and (r), respectively. For example, \( Y_n^{(u)} \) is the set of rooted topologies on \( n \) leaves, and so on. Again, we omit the superscript if the context makes it clear which one is meant.

**Triplet- and quartet-based species tree methods.**—When the gene tree \( \mathcal{G} \) is a rooted tree, we will consider its induced trees on three leaves, which we call triplet trees. However, when the gene tree is unrooted, we will consider its induced trees on sets of four leaves, which we call quartet trees. Thus, for the case of unrooted gene trees, a quartet is a set of four distinct leaves \( X = \{a, b, c, d\} \subseteq L^2 \) and we let \( \mathcal{G}|_X \) be the topology of \( \mathcal{G} \) restricted to \( X \), of which there are three possibilities denoted \( ab|cd, ac|bd, ad|bc \) (where, for the first one, removing the internal path produces two subtrees with leaves \( \{a, b\} \) and \( \{c, d\} \), respectively, etc.).

Let \( Y_n^{(u)} \) be the set of all possible ways of choosing one quartet topology for each quartet in an \( n \)-leaf tree.

We will say that a set of quartet trees is consistent (or compatible) if there is a tree \( T \) that agrees with all the quartet trees (i.e., \( T \) induces each of the quartet trees). Similarly, we will say that a set of triplet trees is compatible or consistent if there is a rooted tree \( T \) that agrees with all the triplet trees.

While most coalescent-based summary methods estimate gene trees and then combine the estimated gene trees, we generalize this approach to allow the method to independently estimate each subset gene tree (triplet gene trees when we assume a strict molecular clock and quartet gene trees, otherwise). These triplet gene trees (or quartet gene trees) are then used to construct the species tree. Thus, when a strict molecular clock is assumed to hold, the first step of such a method would estimate a set of triplet trees for each gene, and the second step would infer the species tree from the different sets of triplet trees. Similarly, when the strict molecular clock is not assumed to hold, the first step would estimate a set of quartet trees for every gene, and then the second step would combine the quartet trees. This approach also allows the method to operate by estimating triplet trees or quartet trees by first estimating a tree on each gene, and then considering the induced triplet trees or induced quartet trees.

We model GTEE in the following way using as an example of the case, where the strict molecular clock is not assumed and so the method estimates quartet trees. We let \( \Delta(Y_n^{(u)}) \) be the set of probability distributions over \( Y_n^{(u)} \). Then a quartet-based GTEE model is a mapping

\[
R : \Delta(Y_n^{(u)}) \rightarrow \Delta(Y_n^{(r)}).
\]

Such a mapping can be equivariant, as above. We denote by \( Y_n^{(u)} \) the collection of quartet topologies reconstructed for gene \( i \).

Note that, in fact, our full gene tree reconstruction model can be seen as a special case of the quartet-based GTEE, which puts positive probability mass only on collections of compatible quartet trees. Therefore, below we work with the more general quartet-based GTEE model.

Also, we extend the definition to the rooted case by considering triplets. For any triplet \( X = \{a, b, c\} \subseteq L^2 \), where \( a, b, c \) are distinct, there are three triplet topologies \( ab|bc, ac|bc, bc|ac \). A triplet-based GTEE model can be defined as above. Again, we denote this case with the superscript \( r \).

**Species Tree Reconstruction**

We seek to solve the following inference problem. We are given \( g \) estimated gene tree topologies (rooted or unrooted) as above. Our goal is to infer the topology of the species tree, \( S \). An inference procedure is said to be statistically consistent if the probability of returning an incorrect reconstruction goes to 0 as \( g \) tends to \( +\infty \).

We consider two cases: where the sequence evolution model for each gene tree is assumed to obey the strict molecular clock and where the strict molecular clock is not assumed to hold.

**Species Tree Estimation in the Presence of a Strict Molecular Clock**

When every gene tree obeys a strict molecular clock (i.e., the substitution rate is constant throughout the tree), we will show the surprising result that for every finite (nonzero) sequence length, it is possible to reconstruct the true species tree with arbitrarily high probability, given a large enough number of gene sequence alignments. The results in this section apply to the Jukes–Cantor model. The methods that achieve this bound are variants on existing methods, and operate in two steps: First triplet trees are computed (independently) for every gene on every three leaves, and then the species tree is computed using these triplet trees. We first describe the latter step in this approach.

**Rooted Triplet Consensus**

We describe a simple method for estimating species trees from \( g \) genes, assuming that we have rooted triplet gene trees for all triplets and for all genes. This method, which we call the Rooted Triplet (RT) consensus, is consistent under the multi-species coalescent model when the strict molecular clock holds and true gene
trees are given as input (see, e.g., Bryant 2003; Ewing et al. 2008; Degnan et al. 2009). Later, we will show that RT is also statistically consistent under a moderate level of GTEE.

Let \( X = \{a, b, c\} \) be a set of three leaves. There are three possible rooted triplet trees on \( X \), which we denote by \( t_0 = ab|c \), \( t_1 = ac|b \), and \( t_2 = bc|a \). By convention, we assume that \( t_0 \) is the triplet tree for these species induced by the species tree.

We first compute the frequency of each triplet tree over all estimated gene trees,

\[
f_{X,0} = \frac{|\{f_X : f_X = t_0\}|}{g}
\]

where \( g \) is the number of genes. We also compute \( f_{X,1} \) and \( f_{X,2} \), the corresponding values for triplet trees \( t_1 \) and \( t_2 \). For each triplet \( X \), we choose the triplet tree with highest frequency (breaking ties arbitrarily); this is called the “dominant triplet tree.”

Recall that a set of triplet trees \( Z = \{z_i\} \) is said to be compatible if there is a tree \( T \in \mathcal{T}_T \), where \( L \) is a leaf set, such that \( T \) agrees with every \( z_i \in Z \). Triplet compatibility is easy to decide (Aho et al. 1981), and when the set \( Z \) includes a single tree on each possible triplet, there is a unique tree on the full set of species that agrees with all the triplet trees. This tree can be constructed in polynomial time using (Aho et al. 1981). Thus, using the approach described above, if the set \( Z \) is compatible, we can return the corresponding tree \( \hat{T} \) in polynomial time, whereas if \( Z \) is not compatible, we return fail. Note that if the dominant triplet tree for each triplet is identical to the species tree for the triplet, then the set \( Z \) will be compatible and the true species tree for the entire set of species will be returned. Finally, if the set \( Z \) is not compatible, we could alternatively apply a triplet-based method for constructing rooted trees from sets of incompatible triplet trees (Snir and Rao 2006).

Topological MP-EST

We also introduce a variant of MP-EST (Liu and Edwards 2010a), and we call this method Topological MP-EST. As with the RT method, Topological MP-EST is statistically consistent under the multi-species coalescent model when the input is a set of true gene trees, but can be shown to be statistically consistent under a moderate level of GTEE. We use the notation above.

Let \( X = \{a, b, c\} \) be a triplet, and assume \( t_0 \) is the triplet tree for the species tree \( S \) on \( X \). It was shown in (Famili and Nei 1988) that, under the multi-species coalescent, for all \( i \), the true gene tree \( \hat{G}_i \), on locus \( i \) satisfies

\[
P_S[\hat{G}_i|X = t_0] = 1 - \frac{2}{3} e^{-\omega_X},
\]

and

\[
P_S[\hat{G}_i|X = t_1] = P_S[\hat{G}_i|X = t_2] = \frac{1}{3} e^{-\omega_X},
\]

where \( \omega_X \) is the length of the internal path in \( S_X \) (in coalescent time units). Then the pseudo-likelihood (where the “pseudo” refers to the fact that we incorrectly assume that the triplets are independent) is

\[
\prod_X g! \frac{g^X}{g^X} \left(1 - \frac{2}{3} e^{-\omega_X}\right)^{g_{X,0}} \left(\frac{1}{3} e^{-\omega_X}\right)^{g_{X,1}} \left(\frac{1}{3} e^{-\omega_X}\right)^{g_{X,2}}.
\]

Topological MP-EST attempts to maximize this pseudo-likelihood over rooted species tree topologies \( \hat{S} \) and internal path lengths \( \omega_X \). Note that unlike (Liu and Edwards 2010a), we do not require the path lengths \( \omega_X \) to be consistent with a choice of branch lengths on \( \hat{S} \). Instead, we set the \( \omega_X \) to their respective maxima for each \( X \) separately, which can be computed as

\[
\hat{\omega}_X = \max \left\{ 0, -\ln \left( \frac{3}{2} (1 - f_X) \right) \right\},
\]

where \( f_X \) is the frequency of the correct triplet under \( \hat{S} \). Note in particular that \( \hat{\omega}_X = 0 \) if \( f_X \leq 1/3 \). This allows for more flexibility in dealing with GTEE. Although this approach ignores correlation between triplet trees, it has computational advantages in that it avoids computing the full likelihood, and hence scales better as the number of genes increases (Liu and Edwards 2010a).

Statistical Consistency under Gene Tree Error

In this section, we establish statistical consistency of RT and Topological MP-EST under GTEE. More specifically, we provide a simple, natural assumption, which we call “triplet-faithfulness” on a gene tree estimation method under which statistical consistency holds. We define triplet-faithfulness in terms of the GTEE model \( R \) for the gene tree estimation method, as follows:

1. \( R \) is triplet faithful if, for all triplets \( X = \{a, b, c\} \), under \( R \) the probability that \( R(G)|X = ab|c \) is greater than the probability of observing either of the other two triplet topologies; that is, if \( G_X = ab|c \) then

\[
P_G[R(G)|X = ab|c] > \max \left\{ P_G[R(G)|X = ac|b], P_G[R(G)|X = bc|a] \right\},
\]

where \( P_G \) denotes the probability measure under \( R(G) \). In other words, for all triplets, the reconstruction method used to estimate gene trees or their triplet topologies favors the correct triplet topology. (This is a special case of the notion of basal centrality in Steel and Rodrigo 2008.)

We will show below that different gene tree estimation methods can satisfy this assumption, but we first show that it implies consistency of RT and Topological MP-EST when gene trees have some estimation error that satisfies triplet-faithfulness.
Triplet Unanimity and Consistency

We establish consistency under triplet-faithful GTEE(0) models for a broad class of summary methods. We say that a reconstruction method has the triplet unanimity property if the following holds.

- For any triplet \( X = \{a, b, c\} \), let \( f_{X,0}, f_{X,1}, f_{X,2} \) be the frequencies of the three triplet topologies on \( X \) in \( (\mathcal{Y})^r \) with \( f_{X,0} \) corresponding to the species tree topology. A species tree reconstruction method has the triplet unanimity property if

\[
f_{X,0} > \max\{f_{X,1}, f_{X,2}\} \quad \forall X
\]

implies that the output \( \hat{S} \) matches the topology of the species tree \( S \).

It is immediate from the definition of RT that it has the triplet unanimity property. We claim that Topological MP-EST also has this property. Indeed, note that for any triplet \( X = \{a, b, c\} \), one can argue without loss of generality that \( S(X) = \{a, b, c\} \).

Theorem 1 (Statistical consistency under triplet unanimity) Assume the true gene trees obey a strict molecular clock, and let \( R \) be a triplet-faithful GTEE(0) model (i.e., a gene tree estimation method that is biased toward the true gene tree). Any species tree reconstruction method having the triplet unanimity property is statistically consistent for the multi-species coalescent when used with gene tree estimation methods with GTEE model \( R \).

Proof: It suffices to show that (2) holds with probability going to 1. We first note that one can argue by a union bound. Therefore, it suffices to show that each term in the sum converges to 0.

The argument will first condition on the gene-specific rate \( \beta_i \) of locus \( i \). We will denote this conditioned probability by \( P_{S,\beta_i} \). Fix a triplet \( X = \{a, b, c\} \) and assume without loss of generality that \( S(X) = \{a, b, c\} \). For a subset \( X' \) of \( X \), we let \( |X'| \) denote the length in coalescent time units of the branch between \( |X'| \) and \( |X| \).

There are two scenarios:

1. Scenario 1: Coalescence in \([a, b]\). First note that if the lineages coming from \( a \) and \( b \) coalesce in \([a, b]\) for locus \( i \) then the true gene tree \( G_i \) shows the correct triplet topology for \( X \). This is an event, which we denote by \( C_{[a, b]} \), of probability \( 1 - e^{-\tau_{ab} \cdot} \).

2. Scenario 2: No coalescence until \([X]\). If this coalescence event does not occur, that is, if \( C_{[a, b]} \) occurs, an event of probability \( e^{-\tau_{ab} \cdot} \), then note that once all lineages enter \([X]\), labels play no role by symmetry. In particular, all gene tree topologies are equally likely. Moreover, recall that we assume the equivariance of the GTEE model \( R \). Hence, by symmetry, we in fact get that all estimated gene tree topologies (restricted to \( X \)) are equally likely as well. That is,

\[
P_{S,\beta_i} \left[ \mathcal{Y} \mid \mathcal{X} = ab \mid c \mid C_{[a, b]} \right] = P_{S,\beta_i} \left[ \mathcal{Y} \mid \mathcal{X} = ac \mid b \mid C_{[a, b]} \right] = P_{S,\beta_i} \left[ \mathcal{Y} \mid \mathcal{X} = bc \mid a \mid C_{[a, b]} \right],
\]

all of which are equal to 1/3.

Hence, overall, we have

\[
P_{S,\beta_i} \left[ \mathcal{Y} \mid \mathcal{X} = ab \mid c \right] = P_{S,\beta_i} \left[ \mathcal{Y} \mid \mathcal{X} = ab \mid c \mid C_{[a, b]} \right] P_{S} \left[ C_{[a, b]} \right] + P_{S,\beta_i} \left[ \mathcal{Y} \mid \mathcal{X} = ab \mid c \mid C_{[a, b]} \right] P_{S} \left[ C_{[a, b]} \right] + P_{S,\beta_i} \left[ \mathcal{Y} \mid \mathcal{X} = ab \mid c \mid C_{[a, b]} \right] \left( 1 - e^{-\tau_{ab} \cdot} \right).
\]

Downloaded from https://academic.oup.com/sysbio/article-abstract/64/4/663/1649531 on 09 April 2018
We first show that site plurality implies triplet-faithfulness.

**Theorem 2 (Site plurality implies triplet faithfulness)**

A triplet estimation method satisfying the site plurality property is necessarily triplet faithful.

**Proof.** Assume \( ab/c \) is the gene tree topology on \( X = \{a, b, c\} \). Let \( 0 < p < q < 3/4 \), where \( p \) is the probability of a change of state at a site along the path from \( [a, b] \) to \( a \) (and hence also to \( b \), since the strict molecular clock holds), and where \( q \) is the probability of a change of state along the path from \( [a, b] \) to \( c \). Then the probability that a site contributes to \( N_{abc}^{ab} \) is

\[
P_{ab} = (1 - p)^2 q + p^2 \left( \frac{1}{3} \right) \left( (1-q) + q \left( \frac{2}{3} \right) \right).
\]

and, similarly, the probability that a site contributes to \( N_{abc}^{ac} \) or \( N_{abc}^{bc} \) is

\[
P_{ac} = (1 - p)(1 - q) p + p (1 - p) q + p^2 \left( 1 - \frac{2}{3} \right).
\]

The difference between these quantities is

\[
P_{ac} - P_{ab} = (1 - p)(1 - q) p + p (1 - p) q - (1 - q) p + p^2 (1 - \frac{2}{3}) > 0.
\]

By coupling (i.e., first generate all sites above with probability \( P_{ab} \) and then add sites contributing to \( N_{abc}^{ac} \) with probability \( P_{ac} - P_{ab} \)) and symmetry, it follows that the probability of the event in (3) is greater than 1/3; this implies triplet-faithfulness.

We give two examples of triplet estimation methods that satisfy the site plurality property—for any finite \( \ell \)—and therefore are triplet-faithful by the previous theorem.

**Basic distance method.**—Consider the following distance-based method.

1. For each pair in \( X \), compute the Hamming distance between the corresponding sequences.
2. If \( xy \) has the lowest Hamming distance value, output \( xy/z \). (Break ties uniformly at random.)

Note that sites contributing to \( N_{abc}^{ab} \) contribute equally to all three distances and sites contributing to \( N_{abc}^{ac} \) do not contribute at all. Therefore, the second step above is equivalent to choosing the pair \( xy \) if it satisfies (3). That proves this method has the site plurality property.

**Maximum likelihood.**—Consider the maximum likelihood estimate. Assume the species tree has topology \( ab/c \) with heights to \( [a, b] \) and \( [a, b, c] \) from the leaves being \( h_{a}^* \) and
h_1$, respectively. To establish the site plurality property, we show that for any fixed heights $h_1 < h_2$, if condition (3) is satisfied, then $ab|c$ has the highest likelihood among the three possible topologies $xy|z$ over $\{a, b, c\}$ with height $h_1$ to $\{x, y\}$ and height $h_2$ to $\{x, y, z\}$. Because this holds for any $h_1, h_2$, it follows that the maximum likelihood gene tree estimate has topology $ab|c$.

Consider estimates $ab|c$, $ac|b$, and $bc|a$ with fixed heights $h_1 < h_2$. Note that the likelihood of sites contributing to $N_{ab}$ and $N_{ac}$ are the same under all three topologies. For $ab|c$, the rest of the log-likelihood is given by

$$N_{ab}^h \log p_{h_1, h_2} + (N_{ac}^h + N_{bc}^h) \log p_{h_1, h_2},$$

where the superscript $h_1, h_2$ indicates that the probabilities in (4) and (5) are computed under this particular choice of heights. For $ac|b$, we have instead

$$N_{ac}^h \log p_{h_1, h_2} + (N_{ab}^h + N_{bc}^h) \log p_{h_1, h_2}.$$  

and similarly for $bc|a$. Note that, under $ac|b$, the quantity $p_{h_1, h_2}$ is the probability that for a given site $a$ and $c$ agree whereas $b$ disagrees, and similarly for $p_{h_1, h_2}$. Taking the difference, we get

$$(N_{ab}^h - N_{ac}^h) \log p_{h_1, h_2}$$

where the logarithm is positive by (6).

We summarize this discussion with the following corollary:

**Corollary 1** Maximum likelihood and Unweighted Pair Group Method with Arithmetic Mean (UPGMA) based on Hamming distances (UPGMA-Hamming) are triplet-faithful gene tree estimation methods when sequence evolution obeys the strict molecular clock under the Jukes–Cantor model. Hence, species trees estimated using the RT or Topological MP-EST methods, applied to triplet gene trees estimated using either maximum likelihood or UPGMA-Hamming will converge to the true species tree as the number of genes increases, for all positive sequence lengths.

**Species Tree Estimation without the Strict Molecular Clock**

When the strict molecular clock cannot be assumed, then a different approach to species tree estimation is needed, since rooted gene trees cannot be estimated with any theoretical guarantees. Instead, we will describe approaches that operate by estimating quartet gene trees rather than triplet gene trees, and then use these quartet gene trees to compute the species tree.

**Quartet Consensus**

We first consider a variant of RT consensus that does not assume a strict molecular clock, which we call Quartet Consensus (QC) (Zhaxybayeva et al. 2006).

Let $X = \{a, b, c, d\}$ be a quartet. There are three possible quartet topologies on $X$, which we denote by $q_0 = ab|cd$, $q_1 = ac|bd$, and $q_2 = ad|bc$. By convention, we assume that $q_0$ is the quartet topology of the species tree. We first compute the frequency of each quartet split over all estimated gene trees,

$$f_{X, g} = \frac{|\{i \in Z|X : q_i = q_0\}|}{g}$$

and similarly for $q_1, q_2$ (where $g$ is the number of genes). For each quartet $X_i$, we choose the quartet topology with highest frequency (breaking ties arbitrarily); the quartet topology with the highest frequency is called the “dominant quartet tree.”

Recall that a set of quartet trees $Z = \{z_i\}$ is said to be compatible if there is a tree $T \in T_n$ with leaf set $L$ such that $T$ agrees with every $z_i \in Z$. Quartet compatibility is, in general, Non-deterministic Polynomial Time (NP)-hard to decide (Steel 1992). However, when the set $Z$ covers all possible quartets, there is a polynomial-time algorithm for deciding compatibility (Bandelt and Dress 1986). In the approach described above, a single quartet tree is chosen for every four species, so that if the set $Z$ is compatible, we can output the corresponding tree $S$ in polynomial time, whereas if $Z$ is not compatible, we return fail. (In the case of incompatible quartets, we could alternatively apply a quartet-based method such as Quartet MaxCut (Snir and Rao 2010).) Furthermore, when the dominant quartet tree is the species tree on each quartet, then the true species tree will be returned.

**Unrooted Topological MP-EST**

We also introduce a variant of Topological MP-EST that does not assume a strict molecular clock, which we call Unrooted Topological MP-EST. We use the notation above.

Let $X = \{a, b, c, d\}$ be a quartet. It was shown in (Allman et al. 2011) that, under the multi-species coalescent, if $q_0$ is the quartet split of the species tree $S$ then, for all $i$, the true gene tree satisfies

$$P_S[G_i | X = q_0] = 1 - \frac{2}{3} e^{-w_X},$$

and

$$P_S[G_i | X = q_1] = P_S[G_i | X = q_2] = \frac{1}{3} e^{-w_X},$$

where $w_X$ is the length of the internal path in $S|X$ (in coalescent time units). Then the pseudo-likelihood
(under the assumption of quartet independence) is

$$\prod_{X} g_{X}! / (g_{X,0})! (g_{X,1})! (g_{X,2})! \left(1 - \frac{2}{3} e^{-\omega_{x}}\right) \left(\frac{1}{3} e^{-\omega_{x}}\right) \left(\frac{1}{3} e^{-\omega_{x}}\right) g_{X,2}!.$$  

Unrooted Topological MP-EST attempts to maximize this pseudo-likelihood over species tree topologies $\hat{S}$ and internal path lengths $\omega_X$. Note that unlike Liu and Edwards (2010a) we do not require the path lengths $\omega_X$ to be consistent with a choice of branch lengths on $\hat{S}$. Instead, we set the $\omega_X$ to their respective maxima, for each $X$ separately, which can be computed as

$$\hat{\omega}_X = \max \left\{ 0, -\ln \left( \frac{3}{2} (1 - f_X) \right) \right\},$$

where $f_X$ is the frequency of the correct quartet split under $\hat{S}$. Also, $\hat{\omega}_X = 0$ if $f_X \leq 1/3$.

Lack of Consistency in the Unrooted Case

We argue in this section that the situation is more significantly complex in the unrooted case.

Quartet-faithfulness is not sufficient.—Let $R$ be an equivariant, quartet-faithful GTEE(u) model. Assume the species tree $S$ has only four leaves $X = \{a, b, c, d\}$ and that its topology is $ab|cd$. We claim that QM may not give a consistent estimator of $S$ in this case. The proof of Theorem 1 relied crucially on two facts: When a deep coalescence occurs, all triplet topologies are equally likely and their distributions as edge-weighted trees are identical; in the absence of a deep coalescence, the gene topology agrees with that of the species tree, giving the species topology a slight advantage overall. The latter is still true in the unrooted case. However, the former is only partly true. Namely, all three topologies are indeed equally likely, but their distributions as weighted trees now differ. For instance, if the branches from $a$ to $\{a, b\}$ and $d$ to $\{c, d\}$ are long and those from $b$ to $\{a, b\}$ and $c$ to $\{c, d\}$ are short, then it is not unreasonable to expect that the gene topology $ab|cd$, being in the Felsenstein zone, may be more likely to lead to errors than the gene topology $ad|bc$. That asymmetry may lead to a failure of QC. Indeed, let $D$ be the event that a deep coalescence occurs on gene tree $G$. Let

$$z_{xy|uv} = \mathbb{P}_G \left[ R(G) \neq xy|uv \mid D \cap \{G = xy|uv\} \right],$$

$$z_{ab|cd} = \mathbb{P}_G \left[ R(G) \neq ab|cd \mid D' \cap \{G = ab|cd\} \right],$$

and $y = \mathbb{P}_G[D]$. Assume it were the case that

$$z_{ac|bd} = z_{ab|cd} = z_{ad|bc} = z_{ad|bc} = e, \quad y = e,$$

for $\varepsilon > 0$ arbitrarily small and that when an estimation error occurs the other two topologies are equally likely to be the output. Note that those values are consistent with quartet-faithfulness and equivariance. Then

$$\mathbb{P}_G[R(G) = ad|bc] = (1 - y) \frac{1}{2} z_{ab|cd} + y \left[ \frac{1}{3} z_{ad|bc} + \frac{1}{6} z_{ab|cd} + \frac{1}{6} z_{ac|bd} \right]$$

$$= \frac{1}{3} + 2 \frac{1}{6} \frac{2}{3} - O(\varepsilon)$$

$$= \frac{5}{9} - O(\varepsilon)$$

$$\geq \frac{1}{2},$$

for $\varepsilon$ small enough. Hence, QC is guaranteed to output the wrong topology in the limit $g \to +\infty$ by the law of large numbers. A similar argument applies to Unrooted Topological MP-EST.

In fact, the situation is even worse: Quartet-faithfulness is likely to fail in general. Consider again the Felsenstein zone example described above. For a single site, maximum likelihood will tend to output the incorrect topology with higher probability than the correct topology (and indeed the branch substitution probabilities can be made so that the probability of outputting the incorrect topology will approach 1) (Tuffley and Steel 1997). Furthermore, it seems possible that for any sequence length the substitution probabilities could be set so that maximum likelihood will tend to output the incorrect topology with higher probability than the true tree.

A Positive Result

As we discussed previously, a consistency result under correctly reconstructed gene trees does not necessarily translate into a consistency result in the presence of GTEE. It is, however, possible to recover a weaker consistency result in the following sense. Suppose gene trees are reconstructed from sequences of length $t$ under the Jukes–Cantor model. It is known (Erdős et al. 1999a, b) that a quartet tree with branch lengths bounded below by $f$ and diameter bounded above by $\Delta$ can be recovered using simple distance-based methods (e.g., the Four Point Method) with probability exceeding $1 - \delta/2$ if

$$\ell \geq C_1 \frac{\varepsilon_0 \Delta}{f^2} \log \frac{3}{\delta},$$

where $C_1$ is a universal constant. When applying this result in the context of ILS, we must be careful to take into account the fact the multi-species coalescent may generate arbitrarily short or long edges. However, as we show below, we can bound the probability that such edges are produced. We make several assumptions. We assume that the diameter of the species tree in coalescent time units is bounded by $T > 0$ and that lineage-specific
Lemma 1. Under the multi-species coalescent, there is a universal constant $C_2 > 0$ such that for any $h > 0$ with probability exceeding $1 - \frac{1}{2}$, a gene tree has no branch length below $\delta(BU)^{-1}/C_2$ and has a diameter bounded above by $BU(T + C_2 \log h^{-1})$.\footnote{Theorem 3}

Proof. Indeed, consider first the probability of producing a short branch. There are $n - 1$ coalescence events, each of which is a race between at most $\binom{2n}{n}$ pairs of lineages, so the probability of producing a branching of length at most $h$ in coalescence time units is bounded above by $(n - 1)(1 - \exp(-\delta(BU)^{-1})) < C_2 h$, where $C_2$ depends on $n$. Taking into account that the lowest mutation rate possible is $(BU)^{-1}$, we have derived an upper bound on the probability of producing a branch shorter than $C_2 h(BU)^{-1}$. On the contrary, consider the probability of producing a large diameter. For any two lineages, the distance in coalescence time units between the corresponding leaves in the species tree is at most $T$. The probability that their coalescence time above their most recent common ancestor is greater than $T - 1$ is at most $\exp(-T)$. Taking into account that the highest mutation rate possible is $BU$, the probability that the diameter of a gene tree is larger than $BU(T + T - 1)$ is at most $\frac{1}{h^2} \exp(-T)$. Choosing $C_2$ appropriately as a function of $C_2$ and $n$ gives the claim. \hfill\vspace{1ex}

Hence, if $\ell \geq C e^{CBU(T + \log h^{-1})^2} \nvo{BU}^{-2} \log h^{-1} \geq C(BU)^2 e^{CBU} h^{-2} + CBU \log h^{-1}$, for some $C > 0$ depending on $n$, then all quartet topologies are correctly reconstructed with probability exceeding $1 - h$. Taking $h$ small enough that $\delta \ll 1 - e^{-f}$, where $f$ is the length of the shortest branch in the species tree in coalescence time units, it follows that QC will succeed with high probability given sufficiently many genes. Indeed, in that case, for each quartet a majority of true gene trees will agree with the species tree and an overwhelming fraction of them will be correctly reconstructed. A similar argument applies to Unrooted Topological MP-EST. Note that this result is significantly weaker than the rooted case in that it requires that for nearly all gene trees, nearly all quartets are correctly reconstructed. (For the rooted case, we only need a bias toward the correct gene tree.)

We summarize this discussion with the following theorem. (This result can be extended to the GTR model under appropriate assumptions on the rate matrices. Details are omitted.)

Theorem 3 Let $S$ be a model species tree, and suppose that genes evolve down $S$ under the multi-species coalescent model. Assume also that sequences evolve down each gene tree under the Jukes-Cantor model, and that quartet gene trees are estimated using a distance-based method. There is a constant $C$ such that if the sequence length $\ell$ satisfies $\ell \geq C(BU)^2 e^{CBU} h^{-2} + CBU \log h^{-1}$, where $f$ is the shortest branch length in the model species tree (in coalescence time units), $\Delta$ is the longest leaf-to-leaf distance in the model species tree (in coalescence time units), and the lineage-specific and gene-specific mutation rates lie, respectively, in the intervals $[U^{-1}, U]$ and $[B^{-1}, B]$, then QC and Unrooted Topological MP-EST applied to the estimated quartet trees will return the true species tree with probability at least $1 - \frac{1}{2}$ from a large enough number of gene trees.


