HHS Public Access

Author manuscript

Bull Math Biol. Author manuscript; available in PMC 2021 July 16.

Published in final edited form as:

Bull Math Biol. ; 82(7): 97. doi:10.1007/s11538-020-00773-4.

Inferring Metric Trees from Weighted Quartets via an Intertaxon Distance

Samaneh Yourdkhani[1], **John A. Rhodes**[1]

[1]Department of Mathematics and Statistics, University of Alaska Fairbanks, Fairbanks 99775, USA

Abstract

A metric phylogenetic tree relating a collection of taxa induces weighted rooted triples and weighted quartets for all subsets of three and four taxa, respectively. New intertaxon distances are defined that can be calculated from these weights, and shown to exactly fit the same tree topology, but with edge weights rescaled by certain factors dependent on the associated split size. These distances are analogs for metric trees of similar ones recently introduced for topological trees that are based on induced unweighted rooted triples and quartets. The distances introduced here lead to new statistically consistent methods of inferring a metric species tree from a collection of topological gene trees generated under the multispecies coalescent model of incomplete lineage sorting. Simulations provide insight into their potential.

Keywords

Phylogenetic tree; Distance; Quartet; Multispecies coalescent; Consensus tree

1 Introduction

We introduce new intertaxon distances that are computed for taxa on an unrooted metric phylogenetic tree based on its displayed rooted triples or quartets. The distances depend upon the *weights*—the lengths of the unique internal edge—of the rooted triples or quartets. These distances differ from the original intertaxon distance on the metric tree, but exactly fit the tree topology, allowing standard distance methods to be used to recover the tree from knowledge of only its weighted rooted triples or quartets. If the rooted triple or quartet data are noisy, so that not all are correct, this distance can still be used to estimate the tree. While the tree estimate will have edge lengths estimating those on a remetrized tree, a simple adjustment gives estimates of the original edge lengths. Thus, these distances lead to new distance-based consensus methodsforobtainingalargemetrictreefromacollectionofweightedrootedtriplesor quartets. In particular, since the weights in coalescent units can be estimated from the distribution of topological gene trees under the standard model of incomplete lineage sorting, these distances can be used in new statistically consistent methods of metric species tree inference from topological gene trees.

Samaneh Yourdkhani syourdkhani@alaska.edu.

This final application is, in fact, our motivation for developing these distances. Statistical inference of a species tree under the multispecies coalescent (MSC) model of incomplete lineage sorting is a fundamental problem in current phylogenetic data analysis. For large datasets (many taxa, with sequences from many loci) that are increasingly common in empirical studies, the simultaneous inference of gene and species trees by Bayesian methods (Liu 2008; Heled and Drummond 2010) may require excessive computation time. Other methods proceed by first inferring gene trees for each locus, and then treating these as data for a second inference of the species tree, by focusing on summaries of the gene trees such as rooted triples, splits, or quartets (Liu et al. 2010; Vachaspati and Warnow 2015; Zhang et al. 2018).

This work continues a thread of developments initiated with several methods of this second sort introduced by Liu and collaborators (Liu et al. 2009; Liu and Yu 2011) for inferring a species tree from a collection of topological gene trees, either rooted or unrooted, under the MSC model. These methods, called STAR and NJ_{st}, proceed by first remetrizing the gene trees in a way that reflects only their topologies, next computing intertaxon distance matrices from each remetrized tree, and then averaging these matrices. Finally, a standard distance method such as Neighbor Joining is used to construct a species tree from this average distance. Despite this seemingly simplistic approach, the methods are statistically consistent under the MSC model (Allman et al. 2013, 2018) and show strong performance in simulation studies (Vachaspati and Warnow 2015). Moreover, they have been shown to be based on the underlying notions of displayed clades and splits on the gene trees (Allman et al. 2013, 2018). A third method, STEAC (Liu et al. 2009), took a similar averaging approach while retaining metric information on the gene trees. Its statistical consistency, however, requires assumptions on the relationship of gene tree metric units (substitution units) to species tree metric units (coalescent units) which may be difficult to justify.

Motivated by the STAR and NJ_{st} algorithms, the RTDC and QDC methods (Rhodes 2019) are based on similar distances defined from displayed topological rooted triples and quartets on gene trees and give statistically consistent inference of topological species trees from gene trees under the MSC. Although the use of the quartet and rooted triple distances results in a slower algorithm than the split or clade approaches of STAR and NJ_{st}, inference with them is more robust to missing taxa on gene trees and gives similar performance to, for instance, the highly developed quartet-based inference software ASTRAL. Moreover, the quartet distance has been generalized to the level-1 network setting (Allman et al. 2019b), playing a key role in the NANUQ method for fast inference of hybridization networks.

While the results presented here are analogs for metric trees of the results for topological trees of (Rhodes 2019), the remetrizations we develop are genuinely new, and not simple extensions of the topological quartet and rooted triple ones. Moreover, since the weights in coalescent units of rooted triples and quartets can be inferred from *topological* gene tree data under the MSC, one can estimate these new intertaxon distances on a species tree from topological gene trees alone. Thus, from the same gene tree data considered in (Rhodes 2019), one obtains not only an estimate of the topology of the species tree, but a metric estimate as well. While the ability to infer a metric species tree is thus similar to STEACs, the approach introduced here crucially uses no *metric* gene tree information, and thus, its

consistency does not depend on any assumptions of the relationship of metric units on gene trees and the species tree. It is thus statistically consistent under much broader assumptions. Although the limited simulation results we present indicate that further work will be necessary to produce algorithms competitive with other approaches, these distances provide new tools for understanding how information on a species tree can be extracted from the gene trees.

Although we position this work in the context of species tree inference, the basic problem of inferring a tree from weighted quartets is not new. Characterizations of those weighted quartet systems that define a metric tree have been given for both binary (Dress and Erdős 2003) and non-binary (Grünewald et al. 2008) trees, insettings where all weights are known exactly. The weighted quartet distanced fined here offers advantages in any setting where there may be noise in the weights, and an exact fit to a single tree is not possible. Then, any of the many methods of fitting a tree to a distance matrix may be applied for an approximate solution. The question of when exact subtree weights uniquely determine a tree was also investigated in (Pachter and Speyer 2004), though the notion of weight in that work is different than is used here.

The remainder of this paper proceeds as follows. After introducing notation and definitions in Sect. 2, the weighted rooted triple metrization and its associated distance are developed in Sect. 3. Section 4 develops the analogs for weighted quartets. Several algorithms using these distances for the inference of a tree from its displayed quartets or a collection of gene trees are formalized in Sect. 5. Finally, Sect. 6 presents some preliminary simulation results and discusses some of the practical issues of using these distance for inference.

Implementations of the quartet versions of the algorithms developed and used in this paper are available in the R package MSCquartets (Allman et al. 2019a).

2 Background and Notation

By a *rooted topological phylogenetic tree* T^r on X, we mean a rooted tree whose root has degree ≥ 2 and all other internal nodes have degree ≥ 3, with leaves bijectively labeled by elements of the finite taxon set X. Directing edges away from the root, we have an ancestral partial order on the nodes, with the root ancestral to all others.

A *rooted metric phylogenetic tree* (T^r, λ^r)on X is a rooted topological tree together with a function λ^r which assigns nonnegative weights, or *edge lengths*, to all edges of T^r. We use T and (T, λ) to denote the unrooted topological and metric species trees obtained from T^r and (T^r, λ^r) in the obvious way, by suppressing the root node if it has degree 2, and undirecting edges.

The *most recent common ancestor* of taxa $x, y \in X$ on a rooted tree T^r is a the minimal node ancestral to both, denoted MRCA(x, y). By the *descendants* of a node v, denoted desc(v), we mean the subset of X labeling leaves that have v as an ancestor.

When considering the *multispecies coalescent model* (MSC) (Pamilo and Nei 1988), we denote its species tree parameter by (σ^r, λ^r). Edge lengths on a species tree are measured in

coalescent units, which are units of time (in generations) inversely scaled by population size, so that the rate of coalescence of two gene lineages in an edge (i.e., population) on the species tree is normalized to 1. Such a parameter determines a probability distribution on rooted and unrooted topological gene trees on X, which we denote as T^r or T. Under the MSC, non-binary topological gene trees have probability 0 even when the species tree is non-binary. Assuming one gene lineage is sampled for each taxon in X, the topological tree σ^r and the edge lengths $\lambda^r(e)$ for all internal edges e on σ^r are identifiable from the distribution of rooted topological gene trees T^r, although lengths of pendant edges on σ^r are not. In fact, σ^r and $\lambda^r(e)$ are identifiable for internal edges e even from the distribution of unrooted topological gene trees T when $|X| \geq 5$. However, if $|X| = 4$, only the unrooted σ and its one internal edge length are identifiable (Allman et al. 2011).

A resolved *rooted triple* is a 3-taxon rooted tree, denoted by $ab|c = ba|c$ where the taxa a, b form a clade. The unresolved rooted triple, a star tree on a, b, c is denoted abc. A rooted tree σ^r or T^r on X *displays* the rooted triples it induces on 3-taxon subsets of X. A *weighted rooted triple* is a pair of a rooted triple together with a weight, a nonnegative real number. We view the weight for a resolved rooted triple as a length for the single internal edge of the triple and allow a weight of zero only if the rooted triple is unresolved. A rooted triple $ab|c$ is said to *separate* the pair a and c, as well as the pair b and c. An unresolved rooted triple does not separate any pairs of taxa on it. The set of rooted triples on X separating taxa a, b is denoted $\mathcal{RT}\ ab$, and the subset of these rooted triples displayed on T^r by $\mathcal{RT}\ ab(T^r)$.

Similarly, a resolved *quartet* is a 4-taxon unrooted tree, denoted by $ab|cd = ba|cd = ab|dc = ba|dc$ where the taxa a,b and c,d form cherries. The unresolved quartet, a star tree on a, b, c, d is denoted $abcd$. An unrooted tree σ or T *displays* the quartets it induces on 4-taxon subsets. A *weighted quartet* is a pair of a quartet together with a weight, a nonnegative real number. We view the weight for a resolved quartet as a length for the single internal edge of the quartet tree and only allow the weight 0 for the unresolved quartet. A quartet $ab|cd$ is said to *separate* the taxon pair a and c, as well as the pairs a, d and b, c and b, d. An unresolved quartet does not separate any pairs of taxa on it. The set of quartets on X separating taxa a, b is denoted as $\mathcal{Q}ab$, and the subset of these quartets displayed on T by $\mathcal{Q}ab(T)$.

Any metric tree (T^r, λ^r) or (T, λ) on X induces a metric d_λ on X, using the sum of edge weights along paths between the taxa. As is well known, however, a metric d on X need not arise from such a weighting. If $d = d_\lambda$ for some λ on T, then we say d is a *tree metric* on T with weighting λ.

For nodes v and w on T, define $P_{v,w} = \{e_1, e_2,...,e_k\}$ to be the path from v to w on T. For a rooted tree T^r, we use the same notation for the set of edges which forms a path from v to w when undirected.

3 Weighted Rooted Triple Metrization of a Rooted Tree

Given a rooted metric tree, we introduce a remetrization of the tree, so that internal edge lengths become a product of their original lengths and an integer factor dependent on the

placement of the edge in the topological tree. Although this introduces no new information, the value of doing this, which will be developed in later section, is to enable an algorithmic approach to inferring a metric tree from its weighted rooted triples, even in the presence of noise. The key theoretical underpinning of this is Theorem 3.1 of this section.

Let (T^r, λ^r) be a rooted metric phylogenetic tree on X. For any vertex v on T^r, denote by $n(v)$ the number of taxa in X which are *not* descendants of v. We remetrize T^r to obtain a new metric tree $(T^r, \tilde{\lambda}^r)$ as follows: First for each internal edge $e = (u, v)$ with u the parent of v, let

$$\tilde{\lambda}^r(e) = \lambda^r(e) \cdot n(v)\,. \tag{1}$$

Then, assign pendant edge lengths in such a way that the tree becomes ultrametric (i.e., all root-to-leaf distance are equal). To do this, we choose any number M greater than the remetrized length of every path of internal edges from the root to any other internal vertex, and to a pendant edge $e = (u, v)$, we assign length

$$\tilde{\lambda}^r(e) = M - \sum_{e \in P_{r,u}} \tilde{\lambda}^r(e) > 0$$

The precise value of M will not matter in what follows, so we assume some choice has been made and fixed. We refer to this remetrization as the *weighted rooted triple metrization*, due to Theorem 3.1.

To further elucidate the need for a choice of M, for $x, y \in X$, let

$$f_{\tilde{\lambda}^r}(x, y) = \sum_{e \in P_{r,\,\mathrm{MRCA}(x,y)}} \tilde{\lambda}^r(e)\,.$$

Then, $-f_{\tilde{\lambda}^r}(x, y)$ is the Gromov product (essentially the Farris transform) (Dress et al. 2007)associated with $d_{\tilde{\lambda}^r}(x, y) = 2(M - f_{\tilde{\lambda}^r}(x, y))$. For $x \neq y$, the Gromov product is independent of the choice of M, but carries all information on the topology of the tree and its internal edge lengths. However, for tree building, it is convenient to pass to a tree metric, which requires a choice of M. Nonetheless, the Gromov product and the tree metric are essentially interchangeable notions.

We now show the intertaxon distance $d_{\tilde{\lambda}^r}$ associated with the weighted rooted triple metrization can also be expressed in terms of information on rooted triple trees induced from T^r. For a fixed tree (T^r, λ^r) on X displaying a rooted triple $xy|z$, let $w(xy|z) = w_{\lambda^r}(xy|z)$ denote the length of the internal edge on the induced metric tree on x, y, z, which we call the *weight* of $xy|z$.

Theorem 3.1

Suppose a rooted metric phylogenetic tree (T^r, λ^r) *is given the rooted triple remetrization,* $(T^r, \tilde{\lambda}^r)$. *Then, for all* $x, y \in X$, $x \neq y$,

$$d_{\tilde{\lambda}^r}(x,y) = 2\left(M - \sum_{xy\,|\,z \text{ on } T^r} w_{\lambda^r}(xy\,|\,z)\right),$$

where the sum is over all $z \in X$ *such that* $xy|z$ *is displayed on* T^r.

Before proving this theorem in full generality, we illustrate it with examples of caterpillar and balanced trees.

Example 3.2—Consider a binary rooted caterpillar tree (T^r, λ^r) on N taxa

$$(\ldots(((a_1,a_2):\lambda_{N-2},a_3):\lambda_{N-3},a_4),\ldots,a_{N-1}):\lambda_1,a_N)$$

with the internal edges of weight $\lambda_1, \lambda_2,\ldots,\lambda_{N-2}$ from the root toward the cherry. Under the rooted triple metrization, for each a_i, a_j, $1 \le i < j \le N$,

$$\begin{aligned} f_{\tilde{\lambda}^r}(a_i,a_j) &= \sum_{e=(v,w)\in P_{r,\mathrm{MRCA}(a_i,a_j)}} \lambda^r(e)\cdot n(w) \\ &= \lambda_1 + 2\lambda_2 + \cdots + (N-j)\lambda_{N-j}. \end{aligned}$$

Also,

$$\begin{aligned} \sum_{a_ia_j\,|\,b \text{ on } T^r} w(a_ia_j\,|\,b) &= \lambda_{N-j} + (\lambda_{N-j} + \lambda_{N-(j+1)}) + \cdots \\ &\quad + (\lambda_{N-j} + \lambda_{N-(j+1)} + \cdots + \lambda_1) \\ &= \lambda_1 + 2\lambda_2 + \cdots + (N-j)\lambda_{N-j}, \end{aligned}$$

where the terms arise from considering, in order, $b = a_{j+1}, a_{j+2},\ldots,a_1$. Thus, $f_{\tilde{\lambda}^r}(a_i,a_j) = \sum_{a_ia_j\,|\,b \text{ on } T^r} w(a_ia_j\,|\,b)$ as Theorem 3.1 shows more generally.

Example 3.3—Let $N = 2^m$ and T^r be a binary rooted balanced tree

$$(\ldots((a_1,a_2),(a_3,a_4)),\ldots,((a_{N-3},a_{N-2}),(a_{N-1},a_N))\ldots)$$

on N taxa. Suppose T^r is given an equidistant metric λ^r where as one moves from the root toward any leaf, the internal edge weights are in order $\lambda_1, \lambda_2,\ldots,\lambda_{m-1}$. Then, edge lengths for $\left(T^r, \tilde{\lambda}^r\right)$ are

$$\tilde{\lambda}_1 = \lambda_1\frac{N}{2},\ \tilde{\lambda}_2 = \lambda_2\frac{3N}{4},\ \tilde{\lambda}_3 = \lambda_3\frac{7N}{8},\ \ldots.$$

Also, if the MRCA(a_i, a_j) is the child vertex of an edge of length λ_k, then

$$f_{\vec{\lambda}}^{r}(a_i, a_j) = \sum_{e=(v,w) \in P_{r,\mathrm{MRCA}(a_i,a_j)}} \lambda^r(e) \cdot n(w) = \frac{N}{2}\lambda_1 + \frac{3N}{4}\lambda_2 + \cdots + N\left(1 - \frac{1}{2^k}\right)\lambda_k .$$

But also

$$\begin{aligned} \sum_{a_i a_j \mid b \text{ on } T^r} w(a_i a_j \mid b) &= \underbrace{(\lambda_k + \cdots + \lambda_1) + \cdots + (\lambda_k + \cdots + \lambda_1)}_{\frac{N}{2} \text{ times}} \\ &\quad + \underbrace{(\lambda_k + \cdots + \lambda_2) + \cdots + (\lambda_k + \cdots + \lambda_2)}_{\frac{N}{4} \text{ times}} \\ &\quad + \cdots + \underbrace{\lambda_k + \cdots + \lambda_k}_{\frac{N}{2^k} \text{ times}} \\ &= \frac{N}{2}\lambda_1 + \left(\frac{N}{2} + \frac{N}{4}\right)\lambda_2 + \cdots \\ &\quad + \left(\frac{N}{2} + \frac{N}{4} + \cdots + \frac{N}{2^k}\right)\lambda_k . \end{aligned}$$

Then, $f_{\vec{\lambda}}^{r}(a_i, a_j) = \Sigma_{a_i a_j \mid b \text{ on } T^r} w(a_i a_j \backslash b)$ as Theorem 3.1 demonstrates more generally.

Proof of Theorem 3.1—With $v = \mathrm{MRCA}(x, y)$ let $r = v_0, v_1, v_2, \ldots, v_n = v$ be the ordered nodes on the path on T^r from the root r to v, as shown in Fig. 1. Let

$$k_i = |\mathrm{desc}(v_i - 1)| - |\mathrm{desc}(v_i)|,$$

the drop in number of descendants from v_{i-1} to v_i. For edge $e_i = (v_{i-1}, v_i)$, let $\lambda_i = \lambda(e_i)$. Then,

$$\sum_{xy \mid z \text{ on } T^r} w_{\lambda}{}^{r}(xy \mid z) = \lambda_n k_n + (\lambda_n + \lambda_{n-1})k_{n-1} + \cdots + (\lambda_n + \lambda_{n-1} + \cdots + \lambda_1)k_1 . \tag{2}$$

For instance, the term $\lambda_n k_n$ on the right side arises because, as can be seen in Fig. 1, there are k_n rooted triple trees $xy|z$, one for each z on the subtree K_n, whose internal edge length is λ_n. While Fig. 1 depicts no polytomies at the v_i, the formula is valid even if polytomies are present.

Rearranging Eq. (2) gives

$$\begin{aligned} \sum_{xy \mid z \text{ on } T^r} w(xy \mid z) &= \lambda_n(k_n + \cdots + k_1) + \lambda_{n-1}(k_{n-1} + \cdots + k_1) + \cdots + \lambda_1 k_1 \\ &= \lambda_n \cdot n(v_n) + \lambda_{n-1} \cdot n(v_{n-1}) + \cdots + \lambda_1 \cdot n(v_1) \\ &= f_{\vec{\lambda}}^{r}(x, y) . \end{aligned}$$

Then, by definition of $d_{\tilde{\lambda}}(x, y)$, we have

$$d_{\tilde{\lambda}}^{r}(x, y) = 2(M - f_{\tilde{\lambda}}^{r}(x, y)) = 2\left(M - \sum_{xy \mid z \text{ on } T^{r}} w(xy \mid z)\right),$$

as claimed.

4 Weighted Quartet Metrization of an Unrooted Tree

For an unrooted metric tree, we define a remetrization similar to that of the last section, using weighted quartets.

Let (T, λ) be an unrooted metric tree on taxa X with $\lambda(e)$ the length of edge e. Each edge e of T determines a split (bipartition) of X, $X = M_e \sqcup N_e$, according to the taxa on the connected components of the graph resulting from deleting e. We remetrize T by assigning to each internal edge e length

$$\tilde{\lambda}(e) = (|M_e| - 1)(|N_e| - 1)\lambda(e),$$

and to pendant edges e length $\tilde{\lambda}(e) = 1$. This gives a new metric tree $(T, \tilde{\lambda})$ which we refer to as having the *weighted quartet metrization*, due to Theorem 4.2. The distance between x and y on the remetrized tree is

$$d_{\tilde{\lambda}}(x, y) = 2 + \sum_{e \in P_{x, y}} (|M_e| - 1)(|N_e| - 1)\lambda(e).$$

We will show this intertaxon distance can also be expressed in terms of information from quartet trees induced from T. As a first step, for a quartet Q, let $E(Q)$ denote the set of edges on the path in T which induces the internal edge of the quartet tree, and $N(e; x, y)$ be the number of quartets $Q \in \mathrm{Q}_{x,y}$, that is, quartets separating x, y, for which $e \in E(Q)$. Then,

$$N(e; x, y) = \sum_{\substack{Q \in \mathcal{Q}_{x, y} \\ e \in E(Q)}} 1. \tag{3}$$

Lemma 4.1

Let T be an unrooted metric phylogenetic tree on taxa X. Then, for all $x, y \in X$, $x \neq y$, and internal edges $e \in P_{x,y}$

$$N(e; x, y) = (|M_e| - 1)(|N_e| - 1) \tag{4}$$

Proof—Let $P_{x,y} = \{e_1,...,e_{n+1}\}$ and $M_i|N_i$ be the split on T associated with e_i. If the path from x to y contains no polytomies, from Fig. 2, we see by Eq. (3) that if k_i denotes the number of taxa on the subtree K_i, then

$$\begin{aligned} N(e_2; x, y) &= k_1k_2 + k_1k_3 + \cdots + k_1k_n = k_1(k_2 + k_3 + \cdots k_n) \\ N(e_3; x, y) &= (k_1k_3 + \cdots + k_1k_n) + (k_2k_3 + \ + \cdots + k_2k_n) \\ &= (k_1 + k_2)(k_3 + \cdots + k_n), \end{aligned}$$

and more generally

$$N(e_i; x, y) = \left(\sum_{j=1}^{i} k_j\right)\left(\sum_{j=i+1}^{k+1} k_j\right) = \left(|M_{e_i}| - 1\right)\left(|N_{e_i}| - 1\right),$$

as claimed. If there are polytomies along the path from x to y, one readily sees the same formula applies. □

For a fixed tree (T, λ) on X displaying a quartet $Q = xy|zv$, let $w(Q) = w_\lambda(Q)$ denote the length of the internal edge on the induced metric tree on x, y, z, v, which we call the *weight* of Q.

Theorem 4.2

Let (T, λ) *be an unrooted, binary metric tree on* X *with* $x, y \in X$. *Then*

$$d_{\tilde{\lambda}}(x, y) = 2 + \sum_{Q \in \mathcal{Q}_{x,y}} w(Q).$$

Proof—By definition of $w(Q)$, we have $w(Q) = \Sigma_{e \in E(Q)} \lambda(e)$. Then,

$$\begin{aligned} 2 + \sum_{Q \in \mathcal{Q}_{x,y}} w(Q) &= 2 + \sum_{Q \in \mathcal{Q}_{x,y}} \sum_{e \in E(Q)} \lambda(e) \\ &= 2 + \sum_{e \in P_{x,y}} \lambda(e) \sum_{\substack{Q \in \mathcal{Q}_{x,y} \\ e \in E(Q)}} 1 \\ &= 2 + \sum_{e \in P_{x,y}} \lambda(e) N(e; x, y) && \text{by equation (3),} \\ &= 2 + \sum_{e \in P_{x,y}} \lambda(e)(|M_e| - 1)(|N_e| - 1) && \text{by Lemma (4.1),} \\ &= d_{\tilde{\lambda}}(x, y). \end{aligned}$$

Example 4.3—The unrooted 8-taxon caterpillar tree

$$(T, l) = (\ldots(((a_1, a_2): \lambda_1, a_3): \lambda_2, a_4), \ldots, a_6): \lambda_5, a_7), a_8),$$

shown in Fig. 3, when remetrized with the quartet metrization $\tilde{\lambda}$ has internal edges of weight

$$(1 \cdot 5)\lambda_1, (2 \cdot 4)\lambda_2, (3 \cdot 3)\lambda_3, (4 \cdot 2)\lambda_4, (5 \cdot 1)\lambda_5,$$

and pendant edges of length 1.

Let $x = a_3$ and $y = a_6$. Then, we have

$$d_{\tilde{\lambda}}(a_3, a_6) = 2 + (2 \cdot 4)\lambda_2 + (3 \cdot 3)\lambda_3 + (4 \cdot 2)\lambda_4 .$$

The 13 quartet trees on T separating a_3 and a_6 are shown in Fig. 4, so

$$\begin{aligned}\sum_{Q \in \mathcal{Q}_{a_3, a_6}} w(Q) &= 2 \cdot \lambda_2 + 1 \cdot \lambda_3 + 2 \cdot \lambda_4 + 2 \cdot (\lambda_2 + \lambda_3) + 2 \cdot (\lambda_3 + \lambda_4) \\ &\quad + 4 \cdot (\lambda_2 + \lambda_3 + \lambda_4) \\ &= (2 \cdot 4)\lambda_2 + (3 \cdot 3)\lambda_3 + (4 \cdot 2)\lambda_4 \\ &= d_{\tilde{\lambda}}(a_3, a_6),\end{aligned}$$

as Theorem 4.2 states.

Example 4.4—Consider an unrooted balanced tree

$$(((a_1, a_2) : \lambda_1,\ (a_3, a_4) : \lambda_2) : \lambda_3,\ ((a_5,\ a_6) : \lambda_4,\ (a_7,\ a_8) : \lambda_5))$$

on 8 taxa as shown in Fig. 5. After remetrization, we have internal edges of weight

$$(1 \cdot 5)\lambda_1, (1 \cdot 5)\lambda_2, (3 \cdot 3)\lambda_3, (1 \cdot 5)\lambda_4, (1 \cdot 5)\lambda_5 .$$

Suppose $x = a_3$ and $y = a_6$. Then,

$$d_{\tilde{\lambda}}(a_3, a_6) = (1 \cdot 5)\lambda_2 + (3 \cdot 3)\lambda_3 + (1 \cdot 5)\lambda_4 .$$

On the other hand, by listing the 13 quartet trees separating a_3 and a_6, we find:

$$\begin{aligned}\sum_{Q \in \mathcal{Q}_{a_3, a_6}} w(Q) &= 2 \cdot \lambda_2 + 4 \cdot \lambda_3 + 2 \cdot \lambda_4 + 2 \cdot (\lambda_2 + \lambda_3) + 2 \cdot (\lambda_3 + \lambda_4) \\ &\quad + 1 \cdot (\lambda_2 + \lambda_3 + \lambda_4) \\ &= (2 + 2 + 1)\lambda_2 + (4 + 2 + 2 + 1)\lambda_3 + (1 + 2 + 1)\lambda_4,\end{aligned}$$

which is equal to $d_{\tilde{\lambda}}(a_3, a_6)$.

5 Weighted Quartet Distance Supertree and Consensus Algorithms

Since, by Theorems 3.1 and 4.2, the pairwise distances between taxa on trees given the rooted triple or quartet remetrizations of the previous sections can be computed from knowing only the weighted rooted triples or weighted quartets displayed on the original tree, they lead to new methods of inferring a large metric tree from that information.

After computing pairwise distances from weighted rooted triples or quartets using the formulas of Theorems 3.1 or 4.2, a standard distance-based tree construction algorithm can

be used to build the remetrized tree. Then, the individual internal edge lengths can be adjusted to remove the multiplier arising from the tree topology in the remetrization. If the tree construction method is robust to some noise, then the presence of a sufficiently small number of erroneous quartets, or sufficiently small errors in the weights, should still allow for construction of an approximation to the original metric tree, with pendant edges weights set to 1.

5.1 Inferring a Tree from Displayed Weighted Quartets

In the quartet case, we present this as a formal algorithm. Let $\mathscr{M}$ denote any method of constructing a metric tree from pairwise distances between taxa. For example, for $\mathscr{M}$, one might choose Neighbor Joining (NJ) (Studier and Keppler 1988) or FastME (Lefort et al. 2015).

Algorithm 5.1—(WQDS/$\mathscr{M}$) Weighted Quartet Distance Supertree with method $\mathscr{M}$

Input: A collection $\mathcal{Q}$ of weighted quartets on taxa in X

1. For each pair $x, y \in X$ of taxa, $x \neq y$, with $\mathcal{Q}_{x,y} \subset \mathcal{Q}$ the subset of weighted quartets separating x and y, define the distance

$$d_{\tilde{\lambda}}(x, y) = 2 + \sum_{Q \in \mathcal{Q}_{x,y}} w(Q).$$

2. Use the distance method $\mathscr{M}$ to build an unrooted metric tree $(T, \tilde{\lambda})$ from $d_{\tilde{\lambda}}$.

3. For each internal edge e on T with associated split $M_e|N_e$, let

$$\lambda(e) = \frac{\tilde{\lambda}(e)}{(|M_e| - 1)(|N_e| - 1)}.$$

For pendant edges e, let $\lambda(e) = 1$.

Output: An unrooted metric tree (T, λ) on X.

The first step of this algorithm, when applied to a set composed of one weighted quartet per choice of 4 taxa in X, has running time $\mathcal{O}(|X|^4)$: One must consider $\binom{|X|}{4}$ quartets, each of which contributes to 4 of the $\binom{|X|}{2}$ sums in that step. If $\mathscr{M}$ is NJ, the second step requires time $\mathcal{O}(|X|^3)$ to obtain a metric tree. By traversing the edges of the tree once, one can compute the M_e, N_e and adjust the edge lengths as in step 3, for an additional time of $\mathcal{O}(|X|)$. Thus, the entire algorithm is accomplished in time $\mathcal{O}(|X|^4)$.

For WQDS to be used, its input of weighted quartet trees must first be obtained. For one genetic locus one might, for example, infer all metric quartet trees on X by standard phylogenetic methods and use the resulting weighted quartets. However, as direct inference of large trees for one locus is already well established and relatively quick, and older quartet methods for this problem are no longer in use, we do not further explore that application.

Instead, we consider a problem of greater current interest: inferring a species tree from a collection of gene trees.

5.2 Inferring a Species Tree from Gene Trees

The standard model for the generation of gene trees from a fixed metric species tree is the *multispecies coalescent model* (MSC) (Pamilo and Nei 1988). The species tree, denoted by σ^r, is rooted with edge weights in *coalescent units*. Coalescent units are obtained from more biologically natural units by inversely scaling the number of generations the edge represents, by the population size, as these cannot be separately identified under the MSC. If the population size is a constant N and the edge represents t generations, the edge weight is simply t/N. If the population varies with time $s \in [0, t]$ along the edge, then the weight is

$$\int_0^t \frac{1}{N(s)} \mathrm{d}s .$$

Under the MSC with one sampled gene lineage per taxon, if the species tree σ^r displays a quartet $ab|cd$ with weight x (the length of the induced quartet tree's internal edge in coalescent units), then the probabilities that a gene tree will display each of the three resolved quartet topologies on these taxa are (Allman et al. 2011)

$$p_{ab\,|\,cd} = 1 - \frac{2}{3}\exp(-x), \quad p_{ac\,|\,bd} = \frac{1}{3}\exp(-x), \quad p_{ad\,|\,bc} = \frac{1}{3}\exp(-x) .$$

If the rooted triple $ab|c$ with weight x is displayed on σ^r, then the same formulas give probabilities of a gene tree displaying rooted triples $ab|c$, $ac|b$, and $bc|a$ respectively (Pamilo and Nei 1988). In particular, since $x > 0$, the quartet or rooted triple with the highest probability of being displayed on a gene tree is the one displayed on the species tree.

This suggests the following algorithm for inferring an unrooted metric species tree from a collection of gene trees under the MSC.

Algorithm 5.2—(WQDC/$\mathscr{M}$) Weighted Quartet Distance Consensus with method $\mathscr{M}$

Input: A collection of n topological gene trees on taxa X

1. For each subset of four taxa $x, y, z, w \in X$, determine the counts of the quartets $xy|zw$, $xz|yw$, and $xw|yz$ displayed on the gene trees.
2. For each subset of four taxa $x, y, z, w \in X$, choose the dominant (i.e., most frequent) quartet as the estimated quartet topology. In the case of a tie, choose from the most frequent uniformly at random. With n_{dom} the number of gene trees displaying the dominant quartet on x, y, z, w, solve the equation

$$1 - \frac{2}{3}e^{-\hat{x}} = \frac{n_{dom}}{n}$$

to find $\hat{x}$ as the estimated weight of the dominant quartet tree.

3. Apply WQDS/$\mathscr{M}$ to the set of $\binom{n}{4}$ estimated weighted dominant quartets.

Output: An unrooted metric tree on X

As discussed in Rhodes (2019), step (1)(a) can be accomplished in time $\mathcal{O}(|X|^4 n)$, with step (2) requiring only time $\mathcal{O}(|X|^4)$. Combined with the computational time for WQDS/$\mathscr{M}$ for $\mathscr{M}$ =NJ shown earlier, the total time is $\mathcal{O}(|X|^4 n)$. Thus, the most time intensive step in the algorithm is tallying the displayed quartets.

Let us say a distance method M of constructing a metric tree from pairwise distances is *well behaved* if (1) when applied to a tree metric returns the unique tree it fits, and (2) is continuous at all tree metrics. The second requirement means that a sufficiently small perturbation in a distance table fitting a binary tree will result in an output of the same binary tree topology, with only small perturbations in the edge weights. Both NJ and Minimum Evolution (ME) are well behaved, though in practice, the heuristic FastME is often used in place of ME.

Theorem 5.3—*Let $\mathscr{M}$ be any well-behaved distance method for tree building. Under the MSC model with one sampled lineage per taxon per gene, on a binary rooted metric species tree (σ^r, λ^r), the output of the WQDC/$\mathscr{M}$ algorithm is a statistically consistent estimator of both the unrooted topological tree σ and the internal edge lengths λ.*

Proof: Consider a collection of n gene trees generated under the MSC on (σ^r, λ^r). Then, for each choice of four taxa x, y, z, w, by the law of large numbers as $n \to \infty$ the probability that the dominant quartet topology matches the quartet displayed on the species tree $\to 1$. Similarly, for any choice of $\epsilon > 0$, the probability that the estimated weight $\hat{x}$ is within ϵ of the quartet weight on the species tree also $\to 1$. Since there are a finite number of sets of 4 taxa, as $n \to \infty$, the probability that all dominant quartet topologies match that on the species tree and all weights are within ϵ of the true value also $\to 1$.

Thus, for any choice of $\epsilon > 0$, with probability $\to 1$ as $n \to \infty$ the computed pairwise quartet distances will be within ϵ of the true values on the species tree with the quartet remetrization. Since $\mathscr{M}$ is well behaved, with probability $\to 1$, it will return the unrooted topology of σ, with internal edge lengths differing from true remetrized values by arbitrarily small amounts. Adjusting the lengths of the internal edges to estimate the original species tree edge lengths involves dividing by a number ≥ 1, so as $n \to \infty$, these estimates can also be made within ϵ of the true values with probability 1. □

It is actually not necessary that all taxa in X are on all gene trees for statistical consistency. As was done in (Rhodes 2019) for the method QDC, one can relax that condition as long as (1) the pattern of missingness of taxa is independent of the gene tree topology, and (2) as the total number of gene trees goes to infinity, so does the number on which each set of 4 taxa appears.

Note that WQDC/$\mathscr{M}$ as presented above does not allow for inference of pendant edge weights on σ. However, if input gene trees have at least 2 samples per taxon, one can infer

those as well, by simply considering an extended species tree obtained by appending two edges of length 0 to each leaf. Similar modifications allow for more samples per taxon.

Remark 5.4: For Weighted Rooted Triple Distance Supertree with method $\mathscr{M}$ (WRTDS/$\mathscr{M}$), one replaces the formulas in steps (1) and (3) of Algorithm 5.2 with similar one arising from Theorem 3.1 and Eq. (1). Note that $\mathscr{M}$ can now be chosen to assume ultrametricity of the distance (e.g., UPGMA), since $\tilde{d}_T$ approximates an ultrametric tree metric. If such an $\mathscr{M}$ is used, then a rooted tree will be returned, and an estimate of both the rooted topology and all its internal edges will be inferred.

Weighted Rooted Triple Distance Consensus with method $\mathscr{M}$ (denoted WRTDC/$\mathscr{M}$) is given by modifying Algorithm 5.2 to count displayed rooted triples and use WRTDS/$\mathscr{M}$.

A consistency result for WRTDC/$\mathscr{M}$ can be shown similarly to Theorem 5.3.

Remark 5.5: In applying WQDC/$\mathscr{M}$ to data, there is one serious practical issue that may need to be addressed. In a finite sample of gene trees, one may find that the dominant quartet for a set of 4 taxa is displayed on every gene tree. Then, solving

$$1 - \frac{2}{3}e^{-x} = 1$$

leads to an estimated weight of ∞ for that quartet. While this correctly indicates the weight should be large, it does not give the finite estimate that is typically needed for applying a tree building method.

Since the MSC does not give expected counts of 100% for one quartet topology for any finite edge weight, this situation can be interpreted as a sign of an insufficient number of gene trees in the data set to properly estimate the weight. One approach to addressing this is to treat counts of $(n, 0, 0)$ for the 3 topologies on a given set of 4 taxa as having dominant count n out of a total of $n + 1$. That is, we reduce the ratio to slightly less than 1. This modification can be viewed as use of a MAP estimate based on a choice of a particular exponential prior for branch lengths (see Sayyari and Mirarab 2016). However, there is no particular motivation for choosing the particular exponential parameter, so the modification remains somewhat *ad hoc*.

This adjustment will result in all infinite weights being replaced by the same finite number. But note that such weights need not result in a good approximation to the desired distance between taxa. A better approach, though one that may not be feasible given practical data collection constraints, is simply to obtain more gene trees so this situation does not occur or restrict to collections of taxa that are closely enough related so that all sets of 4 taxa show some quartet discordance across the gene trees.

As will be shown through simulations in the next section, WQDC/$\mathscr{M}$ may not perform as well as other methods for inferring the topology of the species tree. The reason for this appears to be our inability to obtain accurate estimates of the weight of quartets when they are displayed on all, or almost all, gene trees. While the heuristic described above gives us a

finite estimate which is necessary to have the finite distances between taxa that the algorithm requires, it is unlikely to be very accurate. Even if a handful of gene trees display a quartet other than the dominant one, the estimate of the weight is often not very accurate.

This is not an unusual situation as it often occurs when four taxa are widely placed on a species tree and can occur for taxa whose displayed quartet has only a single edge of the species tree as its internal edge, providing that edge is long in coalescent units. However, simulations suggested to us that a tree inferred by WQDC/$\mathscr{M}$ often did correctly display many correct splits, and those with long edge lengths tended to be correct. That observation is the basis for the following algorithm. It proceeds by using WQDC/$\mathscr{M}$ to pick only one split on the species tree with the largest weight, then dividing the taxa into two groups by this split, and recursively building subtrees on these groups. This process seeks to divide the taxa into smaller groups that will be closer together, so that the poor behavior caused by long edges will not be present in the later stages of the recursion. While it cannot be expected to improve edge length estimates of longer edges, the hope is that the shorter lengths will be estimated well.

Algorithm 5.6—(Recursive WQDC/$\mathscr{M}$) Recursive Weighted Quartet Distance Consensus with method $\mathscr{M}$

Input: A collection of n topological gene trees on taxa X, and positive number L

1. For each subset of four taxa $x, y, z, w \in X$, Determine the counts of the quartets $xy|zw$, $xz|yw$, and $xw|yz$ displayed on the gene trees.
2. If X has 3 or fewer taxa, return the unique unrooted tree on X with all edge lengths 1. Otherwise,
   - **(a)** Apply Steps (2) and (3) of WQDC/$\mathscr{M}$ to the quartet counts obtain an estimated metric species tree $\boldsymbol{\tau}$.
   - **(b)** If all internal edge weights on $\boldsymbol{\tau}$ are less than L, return $\boldsymbol{\tau}$.
   - **(c)** Let $X_0|X_1$ be the split of X associated with the longest edge of $\boldsymbol{\tau}$, and $\ell_{X_0|X_1}$ its length. In the case of a tie, choose the edge uniformly at random from the longest edges.
   - **(d)** Create taxon sets $X_0' = X_0 \cup \{y_1\}$ and $X_1' = X_1 \cup \{y_0\}$, where y_0, y_1 represent "composite taxa" for the split sets X_0, X_1. For each choice of 4 taxa in X_i' compute quartet counts as follows: For quartets containing y_{1-i}, sum over $x \in X_{1-i}$ the counts from Step (1) containing x in place of y_{1-i}. For quartets containing only elements in X_i, retain the quartet counts from Step (1).
   - **(e)** Recursively apply Step (2) to the quartet counts for X_0' and X_1' to obtain metric trees $\boldsymbol{\tau}_0$, $\boldsymbol{\tau}_1$ on X_0', X_1'.

(f) Form a metric tree σ by identifying leaf y_1 on τ_0 with y_0 on τ_1, suppressing that node, and assigning the conjoined edge length $\ell_{X_0|X_1}$. Return σ.

Output: An unrooted metric tree on X

Step (1) requires time $\mathcal{O}(|X|^4 n)$. One application of Step (2) (without the recursive call) on quartet counts for k taxa has time $\mathcal{O}(k^4)$. In the worse case, the split sets have sizes 2, $k-2$ for each recursive call and at every step there is an internal edge weights $\geq L$, leading to time $\mathcal{O}(|X|^4 n + |X|^5)$ for the entire algorithm. However, variations on this algorithm, in which all splits with weights over L in the tree of Step (1), are retained might reduce the typical running time considerably in practical use.

A reasonable choice for the parameter L might be $L = 2$. This corresponds to the quartets defining an edge of length < 2 having an expected frequency of at most $1 - (2/3)\exp(-2) \approx 0.9098$ of the displayed gene quartets matching the species tree quartet.

6 Algorithm Performance in Simulations

Although the algorithms of the last section provide statistically consistent estimators of a species tree from gene trees under the MSC model, their practical performance will be affected by several factors. First, even if gene trees are sampled from the MSC with no error, an algorithm cannot be expected to always infer the underlying species tree from a finite sample of gene trees. Second, if the input gene trees for the algorithm are inferred from sequences that were simulated along the gene trees under some standard substitution model, there is likely to be some inference error in the gene trees due to the finiteness of sequences. Finally, for empirical data, neither the MSC nor the substitution models may exactly describe the true processes, so that there is additional error from model misspecification. Although the performance of phylogenetic inference methods under model misspecification is rarely investigated, simulations can provide insight into the effects of the first two issues.

As an initial, and limited, investigation into the performance of the algorithms of the last section, we present some simulation analysis following the framework of (Rhodes 2019), using the simulated Avian data sets of Bayzid et al. (2015) which were also used in Vachaspati and Warnow (2015). All calculations were performed in R using the ape (Paradis and Schliep 2018) and MSCquartets (Allman et al. 2019a) packages.[1] These data sets for a fixed species tree contain both a sample of gene trees under the MSC, and inferred gene trees from sequences simulated on the sampled gene trees. In addition, there are similar datasets for rescalings of the species tree by factors of 0.5 and 2, to, respectively, increase and decrease the amount of incomplete lineage sorting. For details on the simulation and gene tree inference procedure, see the referenced publications.

To reduce computation time, we pass from the original 48-taxon species tree, to the 30-taxon subtree described in Rhodes (2019). We similarly pass to subtrees of both gene trees sampled under the MSC and subtrees of inferred gene trees. Although these subtrees of

[1] R scripts for the analysis are available at: https://jarhodesuaf.github.io/software.html.

inferred gene trees may not be exactly the trees that would be inferred from the subset of sequences, differences are likely to be small.

Total processing time for a dataset of 1000 gene trees on 30 taxa required about 27–30min for any of the WQDC variant algorithms on a standard laptop computer. The bulk of this time, averaging 27.78min, was for tabulating all quartets displayed on the gene trees, though a more efficient implementation, including parallelism, should reduce this computation time considerably. WQDC required an additional 0.01min, and its recursive variants 1.72min.

We quantify the accuracy of methods in two ways. First, for topological accuracy, we compute the normalized Robinson–Foulds (RF) distance between the true unrooted species tree and the inferred one. The normalization is such that two trees displaying none of the same non-trivial splits will have distance 1, and two binary 30-taxon trees differing by a single NNI move have distance $2/2(30 - 3) = 0.037$.

Second, for metric accuracy, we use a nonstandard variant of a distance of Kuhner and Felsenstein (1994) between the true species tree and the inferred one. For the KF distance as implemented in ape, for each tree, one first forms a vector whose entries correspond to all possible splits of the taxa, with an entry of the length of the edge defined by the split if it is displayed on the tree and 0 otherwise. The Euclidean distance between the vectors for the two trees then gives the distance. The variant we use, denoted KF[x], replaces any vector entry corresponding to a trivial split with 1 and any entry larger than x with x. This treatment of trivial splits is necessary since pendant edge lengths cannot be inferred from this data. The treatment of entries larger than x prevents a split that is displayed on both trees with defining edges of length $\geq x$ but of significantly different sizes from influencing the distance. We use $x = 2$ here, since such long edges on the true species tree give rise to expected quartet counts in which one is large and the others small. These are precisely the counts for which stochastic variation produces large variation in estimated lengths. Note that if two trees differed by splits with edge lengths ≥ 2, then their KF[2] distance would be at least 4. Thus, a KF[2] distance less than 4 indicates the two tree topologies agree on all long edge splits. The choice of 2 here is of course arbitrary, but based on the reasoning given at the end of the last section.

The simulated data sets contain 20 replicates of 1000 sampled and inferred gene trees for each condition, with gene trees inferred from 500 base sequences. In Figs. 6 and 7, we illustrate the mean over the replicates of the distances of inferred species trees from the true one. Results are given for $g = 100$, 400, 700, and 1000 gene trees, by using only the first g gene trees in each simulated collection. We present results of WQDC (Algorithm 5.2) using both the NJ and fastME algorithms for tree building, as well as Recursive WQDC (Algorithm 5.6) using FastME for $L = 2$, 0. For comparison to other methods, we include ASTRAL and QDC, which were already compared for topological inference in Rhodes (2019). Internal edge lengths for trees inferred by these methods, which infer only topological trees, were assigned by methods that use only counts of quartets for sets of four taxa defining those edges, see Zhang et al. (2018) and Allman et al. (2019a) for precise descriptions.

For gene trees sampled from the MSC, with no inference error, Fig. 6 indicates that WQDC, with either distance method, has considerably poorer topological and metric accuracy than the other methods used. While Fig. 7 shows similar results for the methods applied to inferred gene trees, the gap in performance between these methods and others is narrowed. The recursive WQDC, with $L = 0$ or 2, offers a clear improvement over non-recursive WQDC in all situations. This suggests that the source of the poor performance of the non-recursive WQDC is indeed the poor estimation of long edge lengths, as the recursive algorithm operates in such a way that after splits for such edges are put into the tree being inferred, the length of those edges no longer influences future steps. Finally, since there is no substantial difference in the performance of the recursive WQDC for $L = 0$ and $L = 2$, it appears only long edges degrade performance. Since larger values of L reduce running time, this can have an impact for practical use.

When compared to QDC or ASTRAL, the recursive WQDC's performance is usually worse. For topological accuracy, the normalized RF distance is, however, generally less than the 0.037; a single NNI move produces for a 30-taxon tree, so the difference is not great. Interestingly, for metric accuracy, recursive WQDC often matches the best performing algorithm.

Nonetheless, on this, one set of simulations ASTRAL gives the best topological and metric accuracy among all these quartet-based methods. This suggests that if either variant of WQDC is to be useful for empirical inference of species trees, additional development will be needed. We note that while its unweighted analog QDC also is slightly outperformed by ASTRAL, it nonetheless serves as a crucial building block to the NANUQ algorithm for network inference (Allman et al. 2019b), which does have several practical advantages over other network inference methods. There may be similar roles for WQDC.

Acknowledgements

This work was supported by the National Institutes of Health Grant R01 GM117590, awarded under the Joint DMS/NIGMS Initiative to Support Research at the Interface of the Biological and Mathematical Sciences.

References

Allman ES, Degnan JH, Rhodes JA (2011) Identifying the rooted species tree from the distribution of unrooted gene trees under the coalescent. J Math Biol 62(6):833–862 [PubMed: 20652704]

Allman ES, Degnan JH, Rhodes JA (2013) Species tree inference by the STAR method, and generalizations. J Comput Biol 20(1):50–61 [PubMed: 23294273]

Allman ES, Degnan JH, Rhodes JA (2018) Species tree inference from gene splits by unrooted STAR methods. IEEE/ACM Trans Comput Biol Bioinform 15:337–342 [PubMed: 28113601]

Allman ES, Baños H, Mitchell JD, Rhodes JA (2019a) MSCquartets: analyzing gene tree quartets under the multi-species coalescent. R package version 1.0.5. https://CRAN.R-project.org/package=MSCquartets Accessed January 2020

Allman ES, Baños H, Rhodes JA (2019b) NANUQ: a method for inferring species networks from gene trees under the coalescent model. Algorithms Mol Biol 14(24):1–25 [PubMed: 30839948]

Bayzid MS, Mirarab S, Boussau B, Warnow T (2015) Weighted statistical binning: enabling statistically consistent genome-scale phylogenetic analyses. PLoS ONE 10(6):e0129183

Dress AWM, Erdős PL (2003) X-trees and weighted quartet systems. Ann. Comb. 7(2):155–169

Dress A, Huber KT, Moulton V (2007) Some uses of the Farris transform in mathematics and phylogenetics—a review. Ann. Comb. 11(1):1–37

Grünewald S, Huber KT, Moulton V, Semple C (2008) Encoding phylogenetic trees in terms of weighted quartets. J Math Biol 56(4):465–477 [PubMed: 17891538]

Heled J, Drummond AJ (2010) Bayesian inference of species trees from multilocus data. Mol Biol Evol 27(3):570–580 [PubMed: 19906793]

Kuhner MK, Felsenstein J (1994) Simulation comparison of phylogeny algorithms under equal and unequal evolutionary rates. Mol Biol Evol 11:459–468 [PubMed: 8015439]

Lefort V, Desper R, Gascuel O (2015) FastME 2.0: a comprehensive, accurate, and fast distance-based phylogeny inference program. Mol Biol Evol 32(10):2798–2800 [PubMed: 26130081]

Liu L (2008) BEST: Bayesian estimation of species trees under the coalescent model. Bioinformatics 24(21):2542–3 [PubMed: 18799483]

Liu L, Yu L (2011) Estimating species trees from unrooted gene trees. Syst Biol 60:661–667 [PubMed: 21447481]

Liu L, Yu L, Pearl DK, Edwards SV (2009) Estimating species phylogenies using coalescence times among sequences. Syst Biol 58:468–477 [PubMed: 20525601]

Liu L, Yu L, Edwards SV (2010) A maximum pseudo-likelihood approach for estimating species trees under the coalescent model. BMC Evol Biol 10(1):302 [PubMed: 20937096]

Pachter L, Speyer D (2004) Reconstructing trees from subtree weights. Appl Math Lett 17(6):615–621

Pamilo P, Nei M (1988) Relationships between gene trees and species trees. Mol Biol Evol 5(5):568–83 [PubMed: 3193878]

Paradis E, Schliep K (2018) ape 5.0: an environment for modern phylogenetics and evolutionary analyses in R. Bioinformatics 35:526–528

Rhodes JA (2019) Topological metrizations of trees, and new quartet methods of tree inference. IEEE/ACM Trans Comput Biol Bioinfor. 10.1109/TCBB.2019.2917204

Sayyari E, Mirarab S (2016) Fast coalescent-based computation of local branch support from quartet frequencies. Mol Biol Evol 33(7):1654–1668 [PubMed: 27189547]

Studier J, Keppler K (1988) A note on the neighbor-joining algorithm of Saitou and Nei. Mol Biol Evol 5:729–731 [PubMed: 3221794]

Vachaspati P, Warnow T (2015) ASTRID: accurate species trees from internode distances. BMC Genom 16(Suppl 10):S3

Zhang C, Rabiee M, Sayyari E, Mirarab S (2018) ASTRAL-III: polynomial time species tree reconstruction from partially resolved gene trees. BMC Bioinform 19(Suppl 6):15–30

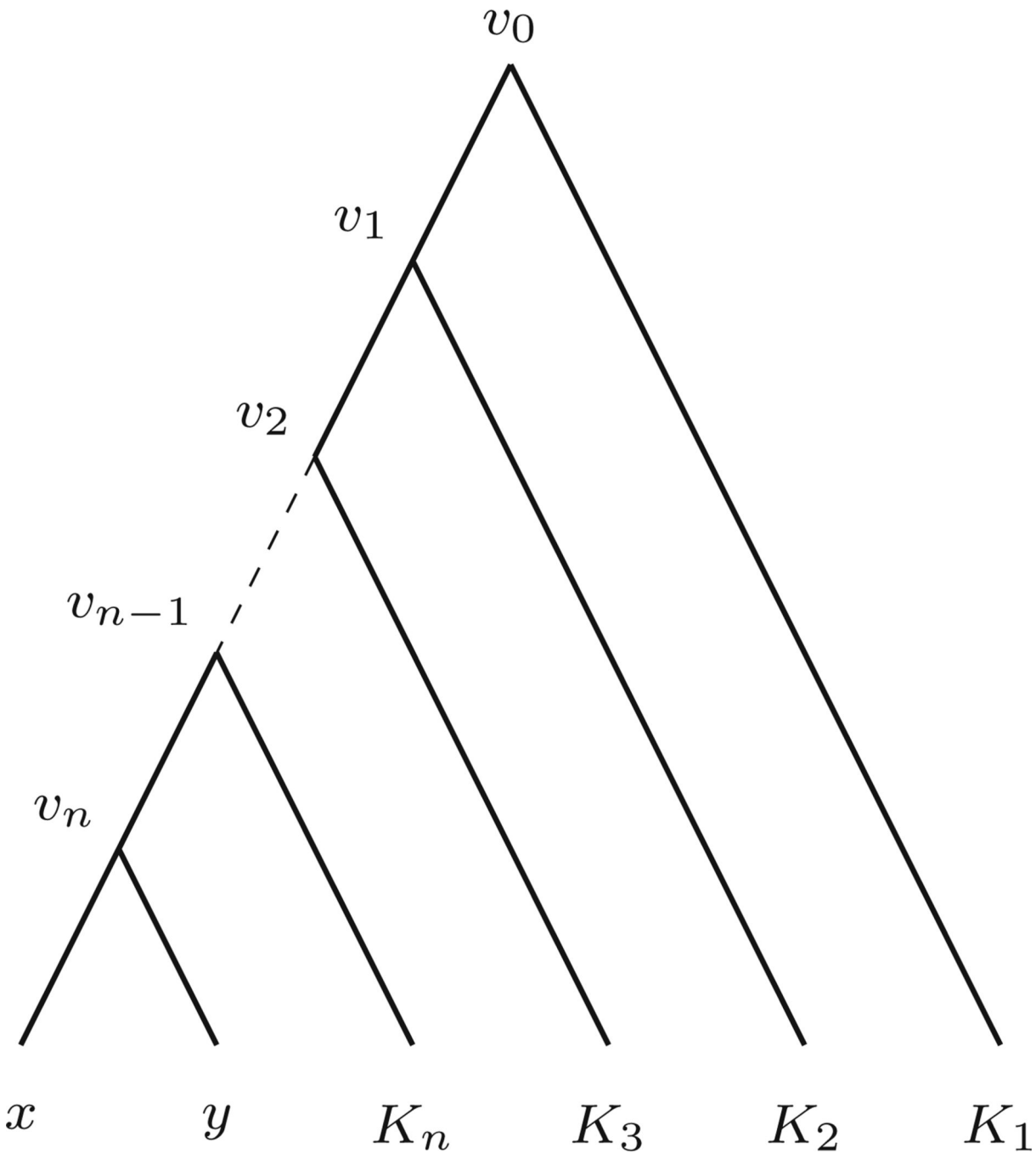

Fig. 1.
An N-taxon binary tree with root v_0 and v_n = MRCA(x, y). The K_i are subtrees, on k_i taxa

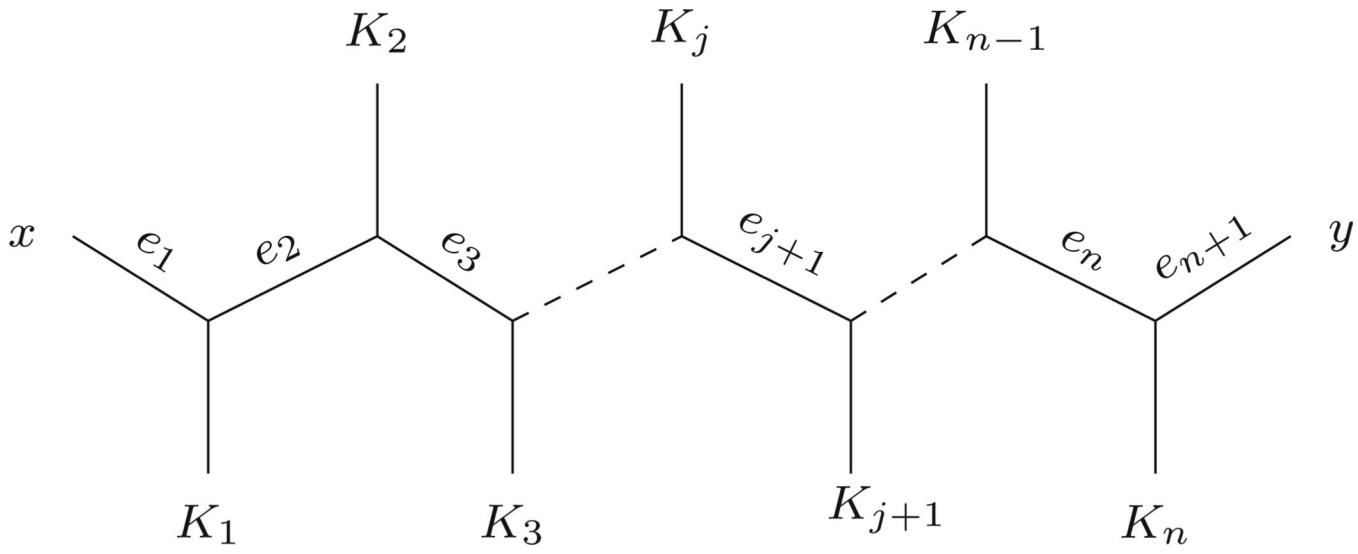

Fig. 2.

The path between taxa x and y on an N-taxon unrooted binary metric tree. The K_i represent subtrees

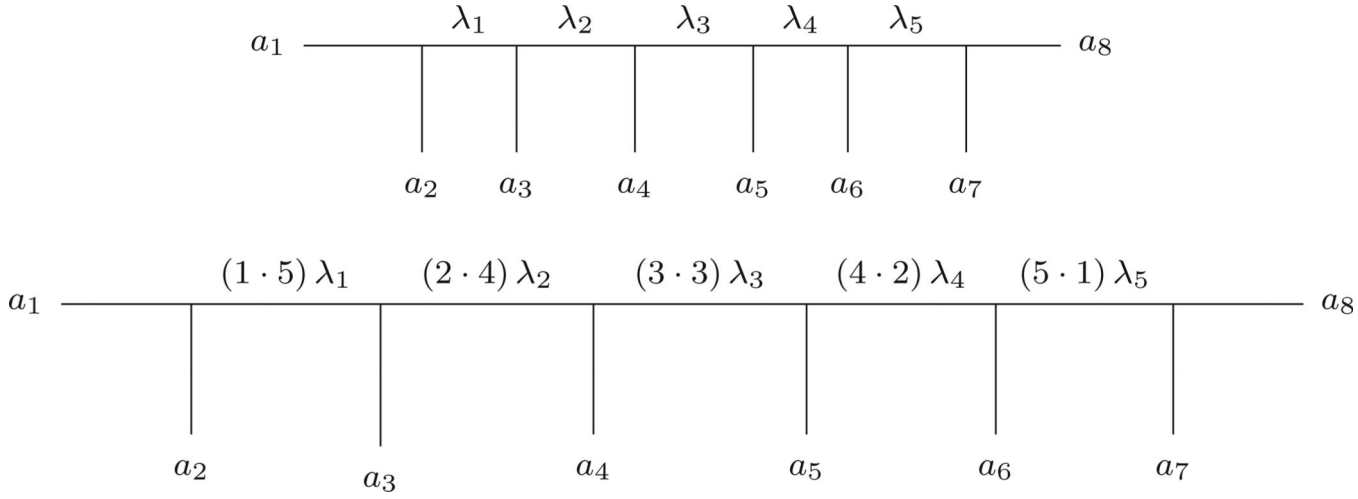

Fig. 3.
An 8-taxon metric caterpillar tree (T, λ) (top) and its quartet remetrization $(T, \tilde{\lambda})$ (bottom)

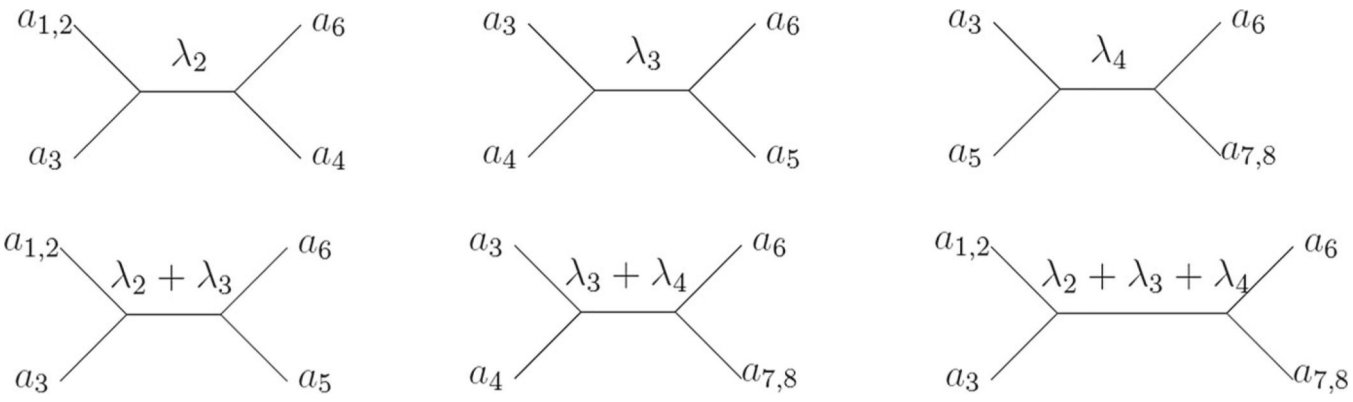

Fig. 4.

The 13 quartet trees on (T, λ) separating a_3 and a_6. Multiple taxa on a leaf represent choices leading to multiple quartet trees

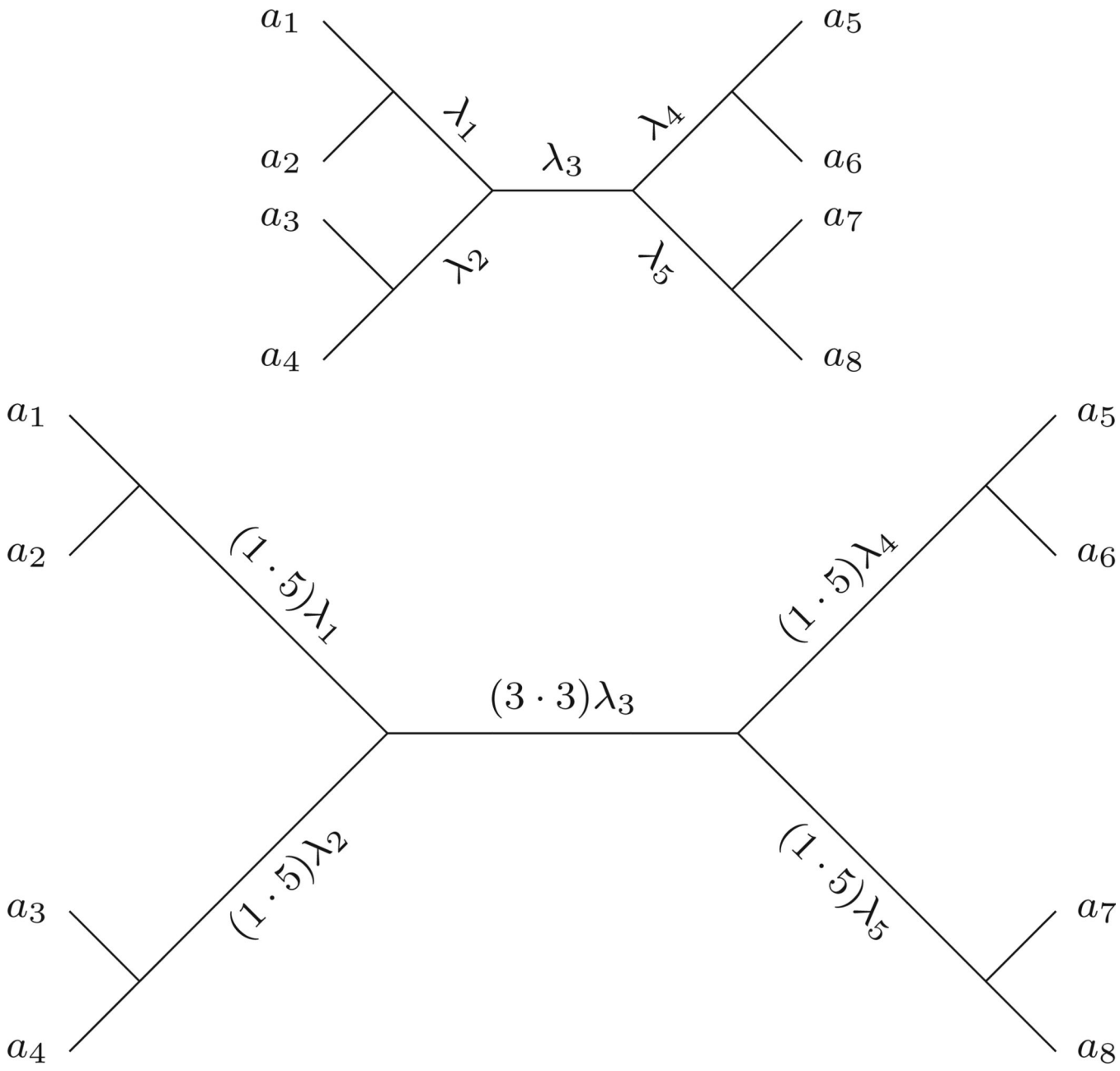

Fig. 5.
An unrooted 8-taxon balanced metric tree, with original edge lengths (top) and quartet remetrization (bottom)

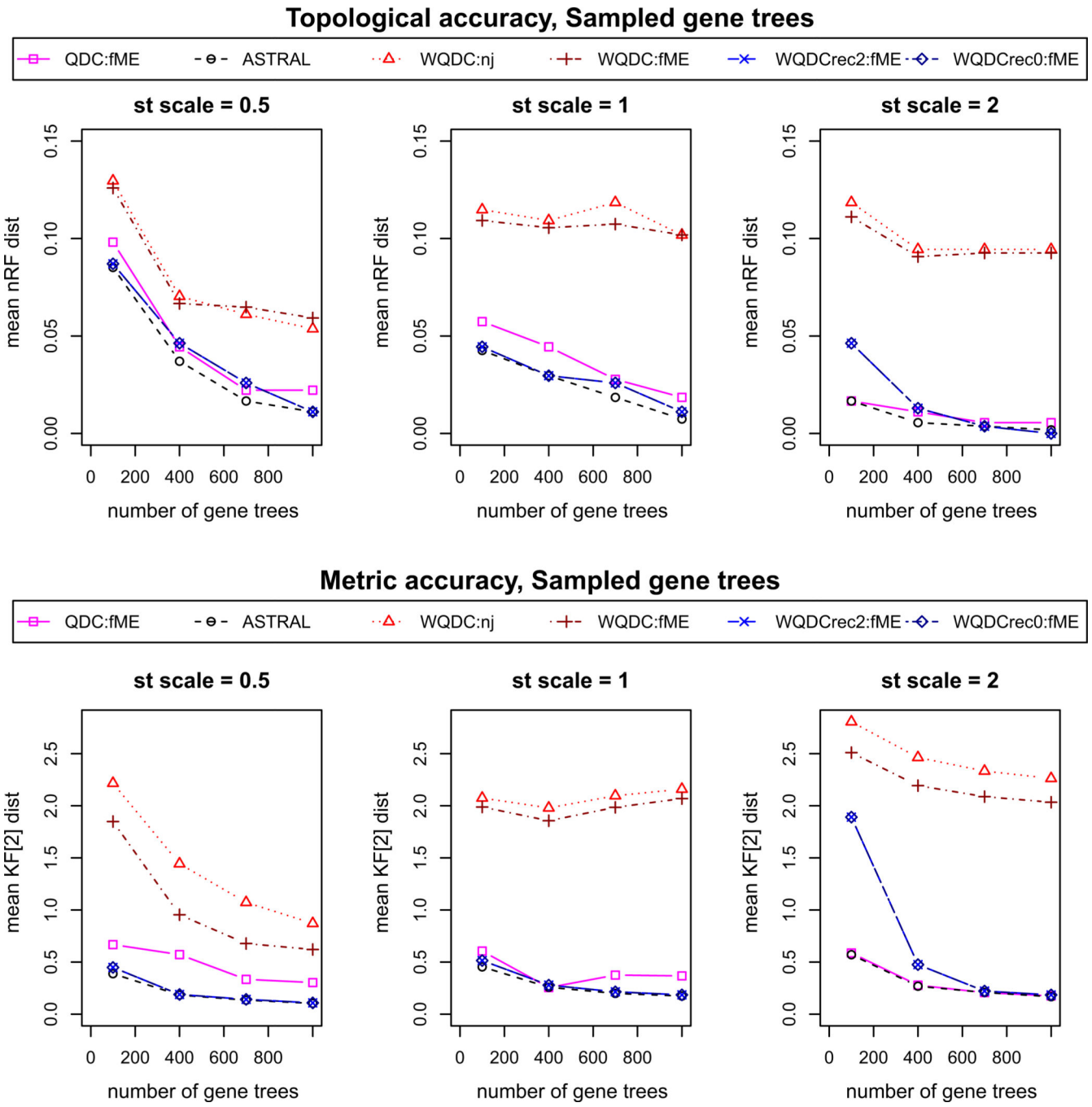

Fig. 6.

Simulation results on accuracy of methods of inference of species trees from gene trees sampled under the MSC

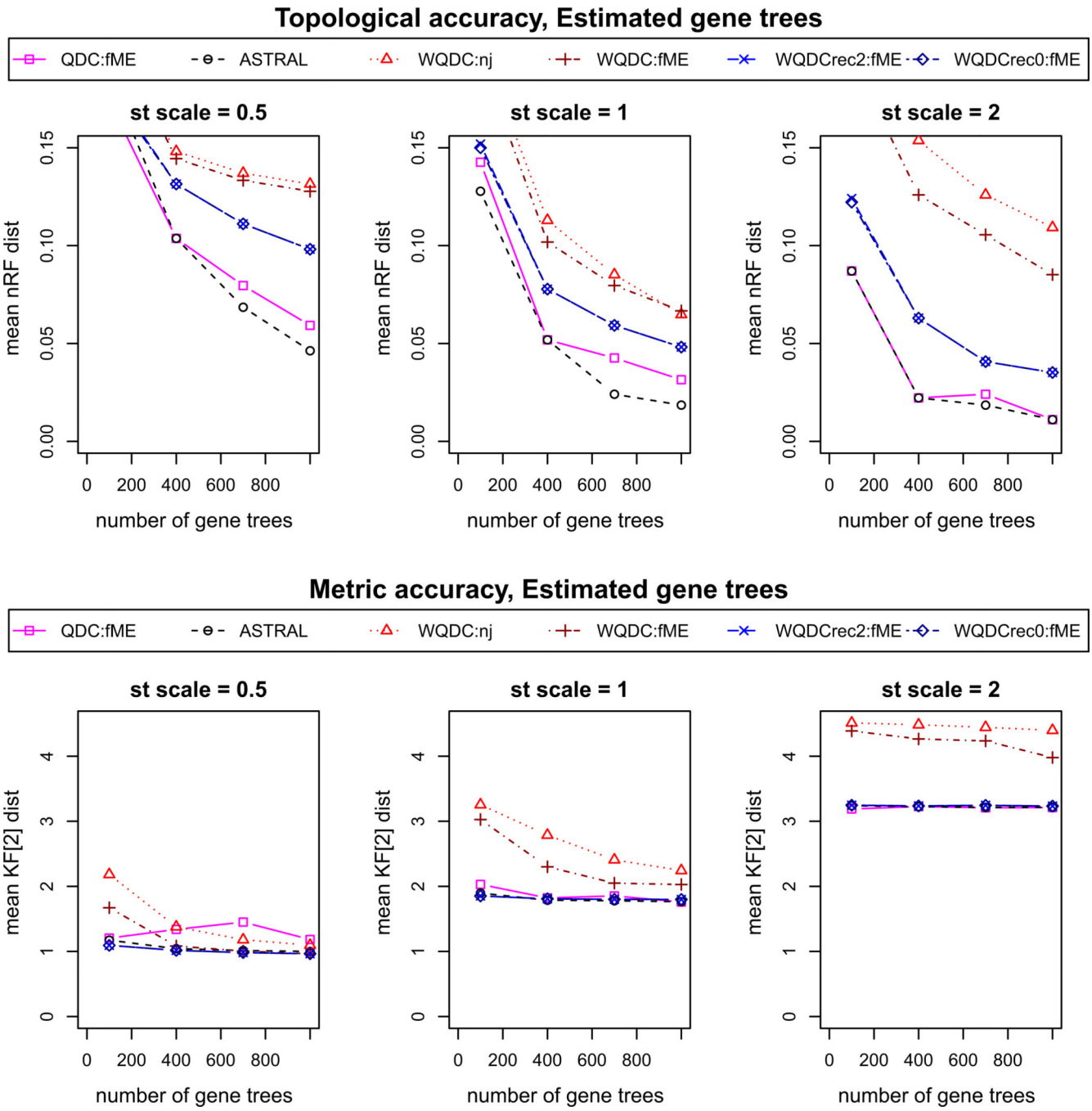

Fig. 7.

Simulation results on accuracy of methods of inference of species trees from gene trees inferred from sequences simulated on trees sampled under the MSC