Ann Inst Stat Math (2017) 69:647–671
DOI 10.1007/s10463-016-0557-x

Distributions of topological tree metrics between a species tree and a gene tree

Jing Xi[1] · **Jin Xie**[2] · **Ruriko Yoshida**[3]

Received: 23 June 2015 / Revised: 27 October 2015 / Published online: 15 March 2016

Abstract In order to conduct a statistical analysis on a given set of phylogenetic gene trees, we often use a distance measure between two trees. In a statistical distance-based method to analyze discordance between gene trees, it is a key to decide "biologically meaningful" and "statistically well-distributed" distance between trees. Thus, in this paper, we study the distributions of the three tree distance metrics: the edge difference, the path difference, and the precise K interval cospeciation distance, between two trees: First, we focus on distributions of the three tree distances between two random unrooted trees with n leaves ($n \geq 4$); and then we focus on the distributions the three tree distances between a fixed rooted species tree with n leaves and a random gene tree with n leaves generated under the coalescent process with the given species tree. We show some theoretical results as well as simulation study on these distributions.

Keywords Coalescent · Phylogenetics · Tree metrics · Tree topologies

✉ Ruriko Yoshida
ruriko.yoshida@uky.edu

Jing Xi
jxi2@ncsu.edu

Jin Xie
jin.xie@uky.edu

[1] Department of Mathematics, North Carolina State University, 2108 SAS Hall, 2311 Stinson Drive, Raleigh, NC 27695, USA

[2] Statistics Department, University of Kentucky, Multidisplinary Science Building, Lexington, KY 40506-0082, USA

[3] Statistics Department, University of Kentucky, 325D Multidisplinary Science Building, Lexington, KY 40506-0082, USA

1 Introduction

A central issue in systematic biology is the reconstruction of populations and species from numerous gene trees with varying levels of discordance (Brito and Edwards 2009; Edwards 2009). While there has been a well-established understanding of the discordant phylogenetic relationships that can exist among independent gene trees drawn from a common species tree (Pamilo and Nei 1988; Takahata 1989; Maddison 1997; Bollback and Huelsenbeck 2009), phylogenetic studies have only recently begun to shift away from single gene or concatenated gene estimates of phylogeny towards these multi-locus approaches (e.g. Carling and Brumfield 2008; Yu et al. 2011; Betancur et al. 2013; Heled and Drummond 2011; Thompson and Kubatko 2013). In order to conduct a statistical analysis on the given set of gene trees, we vectorize each tree, i.e., converting them into a numerical vector format based on a distance matrix or dissimilarity map. These vectorized trees can then be analyzed as points in a multi-dimensional space where the distance between trees increases as they become more dissimilar (Hillis et al. 2005; Semple and Steel 2003; Graham and Kennedy 2010). Such statistical applications that test for incongruence or congruence between two trees using a measurement of dissimilarity between a pair of trees are called distance-based methods (for example, Holmes 2007; Arnaoudova et al. 2010; Weyenberg et al. 2014 are such statistical methods). In a statistical distance-based method to analyze discordance between gene trees, it is a key to decide "biological meaningful" and "statistically well-distributed" distance between trees (Steel and Penny 1993; Coons and Rusinko 2014). Therefore we have studied the distributions of some well-known tree distances between trees. In this paper we focus on three topological tree distances edge difference distance (Williams and Clifford 1971) and precise k-Interval Cospeciation (K-IC) distance (Huggins et al. 2012), and the path difference (Steel and Penny 1993) while the distributions of Robinson–Foulds (RF) distances (Robinson and Foulds 1981) and quartet distances (Brodal et al. 2001) between random trees are very well studied (for example, Steel and Penny 1993).

Here we have conducted simulation studies on these distributions and we have shown theoretical results on the distributions of these tree distances between the *species tree* and *gene trees* which are generated under the coalescent process (Degnan and Salter 2005a).

For the precise K-IC distance between two random trees, Coons and Rusinko (2014) showed that if we take the random trees and compute the distance between them and if we send the number of leaves n of the trees to infinity, then the probability that the distance between two random trees becomes the worst possible distance, that is $(n-3)$, goes to zero while the probability that the RF distance between two random trees becomes the worse possible, that is $2n-6$, goes to one if the trees are selected from a class of trees with a fixed number of cherries (Theorem 8 in Coons and Rusinko 2014). This property is very important to have in terms of applying statistical analysis on the distances of trees. In addition, Steel and Penny (1993) showed some simulation study as well as some theoretical study on the distributions of the RF distance, Quartet distance and path difference distance between random trees with $n = 12$ leaves (see Fig. 6 on Steel and Penny (1993)). A key ingredient of analyzing distributions of these three tree distances between two random trees with n leaves is a simple observation

that the precise K-IC distance between trees is l_∞ norm of two vectorized trees, the path difference distance is l_2 norm of two vectorized trees, and the edge difference distance is l_1 norm of two vectorized trees. First, in this paper, we will show some theoretical results comparing distributions of these tree distances between random trees with n leaves.

A coalescent process is often used to model gene trees given a fixed species tree with n leaves. These theoretical developments have been used to reconstruct species trees from samples of estimated gene trees in practice (Maddison and Knowles 2006; Carstens and Knowles 2007; Edwards et al. 2007; Mossel and Roch 2010; Roy-Choudhury et al. 2008). Rosenberg (2002) studied the distribution of the topological concordance of gene trees and species trees under the coalescent process, Rosenberg (2003) worked on the distributions of monophyly, paraphyly, and polyphyly in a coalescent model, and Degnan and Salter (2005b) studied the distribution of gene trees under the coalescent process. In this paper we focus on the distributions of the edge difference, path difference, and precise K-IC distances between the fixed species tree and gene trees generated under the coalescent process.

This paper is organized as follows. In Sect. 2 we remind readers some definitions. In Sect. 3, we focus on the distributions of these three tree distances between two unrooted random trees. More specifically, in Sect. 3.1, we will show the variance of the distribution of the path difference distance between two random trees with n leaves. In Sects. 3.2 and 3.3 we will compare the means of the distributions of the edge difference and precise K-IC distances between random trees with the mean of the distribution on the path difference distance between them. In Sect. 4, we focus on the distributions of these three different tree distances between a fixed species tree and a gene tree generated from the coalescent process with the species tree. Especially we have computed explicitly the probability that the distribution of any of the three tree distances between a fixed species tree and a gene tree generated under the coalescent process. In Sect. 5, we have shown several simulation studies on the distributions of the three different tree distributions between random trees as well as between a fixed species tree and a gene tree generated from the coalescent. We end with discussions in Sect. 6.

2 Basics and notation

In the subsequent descriptions, let n be the number of leaves (terminal taxa) in a tree. Let $\mathcal{T}_n$ be the space of all possible unrooted trees on n taxa and let $\mathcal{T}_n'$ be the space of all possible rooted trees on n taxa. In this paper we consider only tree metrics between two trees using topological information of the trees, i.e., this tree space does not incorporate branch length information. We use $||\cdot||_p$ to represent the usual l_p norm of a vector, and $|\cdot|$ to indicate the cardinality of a set. A tree distance is a function, $d : \mathcal{T}_n \times \mathcal{T}_n \to \mathbb{R}^+$ that has, at a minimum, the properties $d(r, s) = d(s, r)$ and $d(t, t) = 0$. Many of the methods also require a vectorization function, $v : \mathcal{T}_n \to \mathbb{R}^m$, for some m, which maps phylogenetic trees into Euclidean space. The symmetric difference between two sets is defined as $A \ominus B := (A \backslash B) \cup (B \backslash A)$.

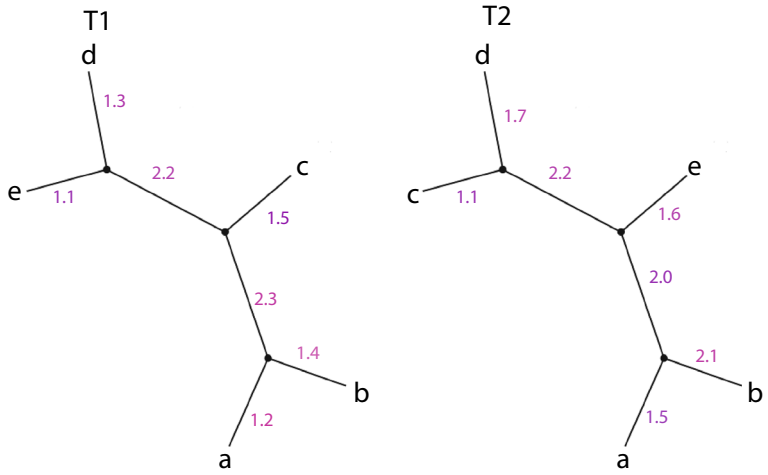

Fig. 1 Example phylogenetic trees: T_1 and T_2. The trees represent proposed most recent common ancestor relationships between 5 taxa, labeled a through e. These trees have branch lengths specified, but not all trees need have such information

Several popular tree distances are squared Euclidean distances as will be demonstrated below.

The dissimilarity map or distance matrix of a tree T is a $n \times n$ symmetric matrix of non-negative real numbers, with zero diagonals and off diagonal elements corresponding to the sum of the branch lengths between pairs of leaves in the tree.

Suppose $v : \mathcal{T}_n \to \mathbb{Z}^{\binom{n}{2}}$ is a function such that the (i, j)th coordinate, where $1 \leq i < j \leq n$, of the $v(T)$ is the number of edges on the unique path between leaves i and j on T.

2.1 Path difference

The RF distance is completely determined by the topologies of the trees, ignoring any edge lengths that may be present. Conversely, the dissimilarity map distance requires that the edge lengths be defined. The *path difference* distance d_P is a distance analogous to the dissimilarity map, but which does not require edge length information.

The calculation of the path difference is identical to the dissimilarity map, except that elements in the distance matrix $D(T)$ are determined by counting the number of edges between the leaves, rather than summing the edge lengths (this is equivalent to the dissimilarity map distance with all edge lengths in the tree set equal to 1). The path difference is studied and compared with the RF distances by Steel and Penny (1993).

Using the lexicographical ordering in the coordinates of the vector, we find that the path difference vectorizations of our example trees in Fig. 1 are

$$v(T_1) = (2, 3, 4, 4, 3, 4, 4, 3, 3, 2),$$
$$v(T_2) = (2, 4, 4, 3, 4, 4, 3, 2, 3, 3).$$

The path difference is therefore, $d_p(T_1, T_2) = ||v(T_1) - v(T_2)||_2 = \sqrt{6}$.

2.2 Edge difference

This tree metric between two trees is defined by Williams and Clifford (1971). Suppose we have two trees $T_1,\ T_2 \in \mathcal{T}_n$. Then the *edge difference* d_e is a distance measure between two trees $T_1,\ T_2 \in \mathcal{T}_n$ such that

$$d_e(T_1, T_2) = ||v(T_1) - v(T_2)||_1.$$

The edge vectorization of any tree is exactly the same as the path difference vectorizations of the tree. The edge difference between trees in Fig. 1 is therefore, $d_e(T_1, T_2) = ||v(T_1) - v(T_2)||_1 = 6$.

2.3 Precise k-interval cospeciaion

The precise k-interval cospeciaion (k-IC) distance d_k is also a distance analogous to the path difference distance, but which uses l_∞ norm instead of l_2 norm. This tree metric was defined by Huggins et al. (2012).

The precise k-IC vectorization of any tree is exactly the same as the path difference vectorizations of the tree. The precise k-IC between trees in Fig. 1 is therefore, $d_k(T_1, T_2) = ||v(T_1) - v(T_2)||_\infty = 1$.

Using the definitions of the tree differences $d_e,\ d_p,\ d_k$ between any two trees $T_1,\ T_2 \in \mathcal{T}_n$ we can immediately have the following remarks.

Remark 1
- The tree differences $d_e,\ d_p,\ d_k$ between any two trees $T_1,\ T_2 \in \mathcal{T}_n$ are tree metrics.
- The tree differences $d_e,\ d_p,\ d_k$ between any two trees $T_1,\ T_2 \in \mathcal{T}_n$ can be computed in $O(n^2)$.
- Many tree metrics such as Nearest-Neighbor-Interchange distance, Subtree-Prune-and-Regraft distance, and Tree-Bisection-and-Regrafting distance are NP-hard (Dasgupta et al. 1997; Hickey et al. 2008; Allen and Steel 2001).

3 Distributions of the three tree metrics between unrooted random trees

In this section we focus on the distributions of the path difference, edge difference and precise K-IC distances between unrooted random trees from $\mathcal{T}_n$.

3.1 Distribution of path difference metric between two trees

Suppose we sampled trees from the uniform distribution over $\mathcal{T}_n$. In this section we consider the distribution of the path difference tree metric d_p between two random trees sampled uniformly from $\mathcal{T}_n$.

Recall that $b(n)$ is the number of binary trees with n labeled leaves. Then we have the following theorems.

Theorem 1 (Theorem 3 from Steel and Penny 1993) *Consider the distribution of d_p^2 under the uniform distribution over $\mathcal{T}_n$. Let $d_{ij}(T)$ for $T \in \mathcal{T}_n$ be the number of edges on the unique path between a leaf i to a leaf j. Then,*

$$\begin{aligned} \mathbf{E}[d_{ij}(T)] &= \alpha(n), \\ \mathbf{V}[d_{ij}(T)] &= 4n - 6 - \alpha(n) - \alpha^2(n), \end{aligned} \tag{1}$$

where $\alpha(n+2) = \frac{2^{2^n}}{\binom{2n}{n}}$ and

$$\mu_p(n) = 2\binom{n}{2}\mathbf{V}[d_{ij}(T)] \tag{2}$$

where $\mu_p(n)$ is the expected value of d_p^2 under the uniform distribution over $\mathcal{T}_n$.

Proof In this paper we only show the proof for $\mu_p(n)$. The rest of the proof for this theorem see Steel and Penny (1993). By definition of d_p^2 we have:

$$d_p^2(T, T') = \| d(T) - d(T') \|_2^2 = \sum_{i<j} [d_{ij}(T) - d_{ij}(T')]^2,$$

where T and T' are two random binary trees. So the mean is:

$$\begin{aligned} \mu_p(n) &= \mathbb{E}[d_p^2(T, T')] = \sum_{T,T'} \Pr(T)\Pr(T') d_p^2(T, T') \\ &= \sum_{T,T'} \frac{1}{b(n)^2} \sum_{i<j} [d_{ij}(T) - d_{ij}(T')]^2 \\ &= \frac{1}{b(n)^2} \sum_{T,T'} \sum_{i<j} [d_{ij}(T)^2 + d_{ij}(T')^2 - 2 d_{ij}(T) d_{ij}(T')] \\ &= \frac{1}{b(n)^2} \sum_{i<j} \left[\sum_{T,T'} d_{ij}(T)^2 + \sum_{T,T'} d_{ij}(T')^2 - 2 \sum_{T,T'} d_{ij}(T) d_{ij}(T') \right] \\ &= \frac{1}{b(n)^2} \sum_{i<j} \left[\sum_{T'} \left(\sum_{T} d_{ij}(T)^2 \right) + \sum_{T} \left(\sum_{T'} d_{ij}(T')^2 \right) \right. \\ &\quad \left. -2 \sum_{T} d_{ij}(T) \left(\sum_{T'} d_{ij}(T') \right) \right] \\ &= \frac{1}{b(n)^2} \sum_{i<j} \left[2b(n) \sum_{T} d_{ij}(T)^2 - 2 \left(\sum_{T} d_{ij}(T) \right)^2 \right]. \end{aligned}$$

Notice that $\sum_T f(d_{ij}(T))$ does not depend on the selection of i and j because of the symmetry of labeling (it is easy to prove by contradiction and switching the labels). Therefore $\sum_T f(d_{ij}(T)) = \sum_T f(d_{kl}(T))$ with $i < j, k < l$, and thus we have:

$$\begin{aligned}
\mu_p(n) &= \frac{2}{b(n)^2}\binom{n}{2}\left[b(n)\sum_T d_{ij}(T)^2 - \left(\sum_T d_{ij}(T)\right)^2\right] \\
&= 2\binom{n}{2}\left[\sum_T \frac{d_{ij}(T)^2}{b(n)} - \left(\sum_T \frac{d_{ij}(T)}{b(n)}\right)^2\right] \\
&= 2\binom{n}{2}\left[\sum_T d_{ij}(T)^2 \Pr(T) - \left(\sum_T d_{ij}(T)\Pr(T)\right)^2\right] \\
&= 2\binom{n}{2}\left(\mathbb{E}\left[d_{ij}(T)^2\right] - \mathbb{E}\left[d_{ij}(T)\right]^2\right) = 2\binom{n}{2} Var(d_{ij}(T))
\end{aligned}$$

with any selection of i and j. □

Theorem 2 *$\sigma_p^2(n)$, the variance of d_p^2, is*

$$\begin{aligned}
\sigma_p^2(n) = \frac{1}{b(n)^2}\Bigg\{ & \sum_{T,T'}\left[\sum_{i<j} d_{ij}(T)^2\right]^2 + \sum_{T,T'}\left[\sum_{i<j} d_{ij}(T')^2\right]^2 + 4\sum_{T,T'}\left[\sum_{i<j} d_{ij}(T)d_{ij}(T')\right]^2 \\
& +2\left\{\binom{n}{2} b(n)[4n-6-\alpha(n)]\right\}^2 \\
& -8b(n)\alpha(n)\sum_T\left[\sum_{i<j} d_{ij}(T)^2\right]\left[\sum_{i<j} d_{ij}(T)\right] \Bigg\} \\
& -4\left[\tbinom{n}{2}\mathbf{V}[d_{ij}(T)]\right]^2.
\end{aligned}$$

Proof Since $\sigma_p^2(n) = Var(d_p^2) = \mathbb{E}[d_p^4] - \mu_p(n)^2$, where the explicit formula of $\mu_p(n)$ is known, we only need to consider only $\mathbb{E}[d_p^4]$:

$$\begin{aligned}
\mathbb{E}[d_p^4(T,T')] &= \sum_{T,T'} \Pr(T)\Pr(T')[d_p^2(T,T')]^2 \\
&= \sum_{T,T'} \frac{1}{b(n)^2}\left(\sum_{i<j}[d_{ij}(T) - d_{ij}(T')]^2\right)^2 \\
&= \frac{1}{b(n)^2}\sum_{T,T'}\left[\sum_{i<j} d_{ij}(T)^2 + \sum_{i<j} d_{ij}(T')^2 - 2\sum_{i<j} d_{ij}(T)d_{ij}(T')\right]^2
\end{aligned}$$

$$
\begin{aligned}
&= \frac{1}{b(n)^2}\left\{\begin{array}{l}
\sum_{T,T'}\left[\sum_{i<j} d_{ij}(T)^2\right]^2 + \sum_{T,T'}\left[\sum_{i<j} d_{ij}(T')^2\right]^2 \\
\quad +4\sum_{T,T'}\left[\sum_{i<j} d_{ij}(T)d_{ij}(T')\right]^2 \\
\quad +2\sum_{T,T'}\left[\sum_{i<j} d_{ij}(T)^2\right]\left[\sum_{i<j} d_{ij}(T')^2\right] \\
\quad -4\sum_{T,T'}\left[\sum_{i<j} d_{ij}(T)^2\right]\left[\sum_{i<j} d_{ij}(T)d_{ij}(T')\right] \\
\quad -4\sum_{T,T'}\left[\sum_{i<j} d_{ij}(T')^2\right]\left[\sum_{i<j} d_{ij}(T)d_{ij}(T')\right]
\end{array}\right\} \\
&= \frac{1}{b(n)^2}\left\{\begin{array}{l}
\sum_{T,T'}\left[\sum_{i<j} d_{ij}(T)^2\right]^2 + \sum_{T,T'}\left[\sum_{i<j} d_{ij}(T')^2\right]^2 \\
\quad +4\sum_{T,T'}\left[\sum_{i<j} d_{ij}(T)d_{ij}(T')\right]^2 \\
\quad +2\sum_{T,T'}\left[\sum_{i<j} d_{ij}(T)^2\right]\left[\sum_{i<j} d_{ij}(T')^2\right] \\
\quad -8\sum_{T,T'}\left[\sum_{i<j} d_{ij}(T)^2\right]\left[\sum_{i<j} d_{ij}(T)d_{ij}(T')\right]
\end{array}\right\}.
\end{aligned}
$$

In this equation, two terms can be simplified as:

$$
\begin{aligned}
\sum_{T,T'}\left[\sum_{i<j} d_{ij}(T)^2\right]\left[\sum_{i<j} d_{ij}(T')^2\right] &= \left[\sum_{T}\sum_{i<j} d_{ij}(T)^2\right]\left[\sum_{T'}\sum_{i<j} d_{ij}(T')^2\right] \\
&= \left[\binom{n}{2}\sum_{T} d_{ij}(T)^2\right]^2 \\
&= \left\{\binom{n}{2} b(n)\mathbb{E}[d_{ij}(T)^2]\right\}^2 \\
&= \left\{\binom{n}{2} b(n)[4n-6-\alpha(n)]\right\}^2.
\end{aligned}
$$

$$\sum_{T,T'}\left[\sum_{i<j} d_{ij}(T)^2\right]\left[\sum_{i<j} d_{ij}(T)d_{ij}(T')\right] = \sum_{T}\left[\sum_{i<j} d_{ij}(T)^2\right] \times \left[\sum_{T'}\sum_{i<j} d_{ij}(T)d_{ij}(T')\right]$$
$$= \sum_{T}\left[\sum_{i<j} d_{ij}(T)^2\right] \times \left[\sum_{i<j} d_{ij}(T)b(n)\mathbb{E}[d_{ij}(T)]\right]$$
$$= b(n)\alpha(n)\sum_{T}\left[\sum_{i<j} d_{ij}(T)^2\right] \times \left[\sum_{i<j} d_{ij}(T)\right].$$

□

3.2 Distribution of the edge difference metric between two trees

Theorem 3 *Consider the distribution of d_e under the uniform distribution over $\mathcal{T}_n$. Then, using the relation between l_p norm and l_q norms where $0 < q < p$ such that $||x||_p \leq ||x||_q \leq m^{(\frac{1}{q}-\frac{1}{p})}$, we have the following theorem:*

$$\sqrt{2\binom{n}{2}\left(4n-6-\alpha(n)-\alpha^2(n)\right)} \leq \mu_e(n) \leq \binom{n}{2}\sqrt{2\left(4n-6-\alpha(n)-\alpha^2(n)\right)} \tag{3}$$

where $\mu_e(n)$ is the expected value of d_e under the uniform distribution over $\mathcal{T}_n$.

Remark 2 Let $B(x) = \sum_{n>0} \frac{b(n+1)}{n!}x^n$ be an exponential generating function for the number of planted binary trees, $b(n+1)$, with n labeled non-root leaves (or the number of rooted binary trees with n leaves). Let

$$F(x, y) = yB(x) + y^2B(x) + \cdots = \frac{1}{[1 - yB(x)]} - 1$$

be the exponential generating function for the number of ordered forests consisting of a given number of rooted trees (marked by y) and a given number of leaves (marked by x). Then for a fixed pair of distinct leaves i and j (we can set $i = 1$ and $j = 2$), we have

$$\sum_{T\in\mathcal{T}_n}\sum_{T'\in\mathcal{T}_n}|d_{ij}(T)-d_{ij}(T')| = \sum_{r=2}^{n-1}[y^r][x^{n-2}]yF(x,y) \times\left(\sum_{r'=2}^{n-1}|r-r'|[y^{r'}][x^{n-2}]yF(x,y)\right),$$

where $[x^k][y^{k'}]f(x,y)$ denotes the coefficient of $x^k\cdot y^{k'}$ in the function $f(x,y)$.

3.3 Distribution of the precise k-IC tree metric between two trees

Now we consider the distribution of d_k under the uniform distribution over $\mathcal{T}_n$. Then, using the relation between l_p norm and l_q norms where $0<q<p$ such that $||x||_p \le ||x||_q \le m^{(\frac{1}{q}-\frac{1}{p})}$, we have the following theorem:

Theorem 4 *Consider the distribution of d_k under the uniform distribution over $\mathcal{T}_n$. Then,*

$$\sqrt{2\left(4n-6-\alpha(n)-\alpha^2(n)\right)} \le \mu_k(n) \le \sqrt{2\binom{n}{2}\left(4n-6-\alpha(n)-\alpha^2(n)\right)} \quad (4)$$

where $\mu_k(n)$ is the expected value of d_k under the uniform distribution over $\mathcal{T}_n$.

Remark 3 Using the same relation above, we can use $\mu_k(n)$ as an upper bound for $\sqrt{\mu_p(n)}$ and $\mu_e(n)$, that is

$$\begin{aligned}\sqrt{\mu_p(n)} &\le \sqrt{\tbinom{n}{2}}\mu_k(n)\\ \mu_e(n) &\le \tbinom{n}{2}\mu_k(n).\end{aligned}$$

4 Species tree and gene tree under the coalescent

Let $\mathcal{T}'_n$ be the space of rooted trees with n leaves. Note that $\mathcal{T}'_n=\mathcal{T}_{n+1}$. In this section we consider the distances between a species tree and a gene tree under the coalescent given the species tree. First we consider the following two lemmas from Coons and Rusinko (2014).

Lemma 1 (Lemma 1 from Coons and Rusinko 2014) *For any two trees $T_1,\ T_2\in\mathcal{T}'_n$, $d_k(T_1,\ T_2)\le(n-2)$.*

A caterpillar tree is any unrooted binary phylogenetic tree which reduces to the path if we delete all edges attached to a leaf and all leaves (see Fig. 3 for an example).

Lemma 2 (Corollary 1 from Coons and Rusinko 2014) *If $d_k(T_1,\ T_2)=(n-2)$ for $T_1,\ T_2\in\mathcal{T}'_n$, then T_1 or T_2 is a caterpillar tree.*

Coons and Rusinko (2014) considered unrooted trees in $\mathcal{T}_n$. In the case of unrooted trees in $\mathcal{T}_n$, we have the bound $(n-3)$ in Lemmas 1 and 2. But in this section we

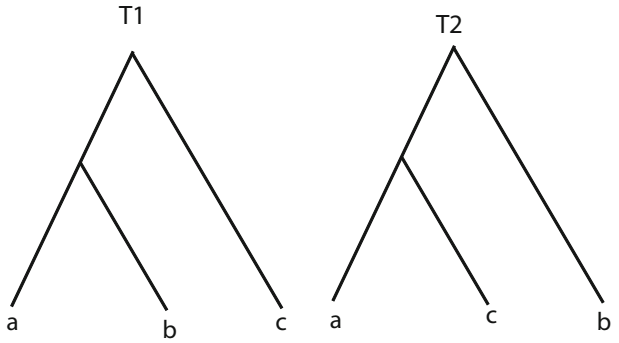

Fig. 2 Example of phylogenetic rooted trees: T_1 and T_2. The trees represent proposed most recent common ancestor relationships between 3 taxa, labeled a through c

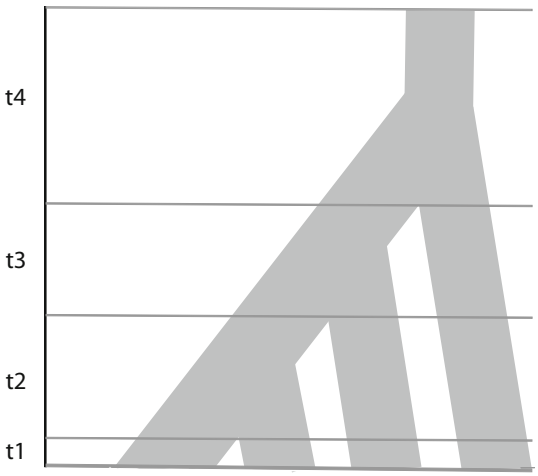

Fig. 3 The caterpillar species tree T_s with $n = 4$

consider $\mathcal{T}_n'$, the space of rooted trees and using the fact that $\mathcal{T}_n' = \mathcal{T}_{n+1}$, thus we have the bound $((n+1)-3) = (n-2)$. For example, if we consider T_1 and T_2 in $\mathcal{T}_n'$ as seen in Fig. 2, then $d_k(T_1,\ T_2) = ||(2,3,3) - (3,2,3)||_\infty = (3-2) = 1$.

Thus, a caterpillar tree is a special case, so we consider that the species tree $T_s \in \mathcal{T}_n'$ be a caterpillar tree. In this section we also consider a sample size of individuals from each species is one and each species has the same effective population size N_e. Let t_i be a time interval in the coalescent time unit between the $(i-1)$th event when two species are coalesced to the ith event when two species are coalesced (see Fig. 3).

Let $T_s \in \mathcal{T}_n'$ be a caterpillar tree. Now we consider the probability that $T_s \in \mathcal{T}_n'$ and a gene tree T_g generated by the coalescent given the species tree T_s have the same tree topology.

Let $g_{ij}(t)$ be the probability that i lineages derive from j lineages that existed $t > 0$ coalescent time units in the past such that

$$g_{ij}(t) = \sum_{k=j}^{i} \exp\left(\frac{-k(k-1)t}{2}\right) \frac{(2k-1)(-1)^{k-j} j_{(k-1)} i_{[k]}}{j!(k-j)! i_{(k)}},$$

where $a_{(k)} = a(a+1)\dots(a+k-1)$ for $k \geq 1$ with $a_{(0)} = 1$; and $a_{[k]} = a(a-1)\dots(a-k+1)$ for $k \geq 1$ with $a_{(0)} = 1$ (Takahata 1989; Takahata and Nei 1990; Tavaré 1984). $g_{ij}(t) = 0$ except with $1 \leq j \leq i$.

Remark 4 If t is a scale of coalescent time units then t can be written as $t = \frac{t'}{N_e}$ where t' is the number of generation and N_e is a population size. We assume that the size of an ancestral species is the sum of the sizes of its descendants so that the scaling of time would be different before and after the divergence of the ancestor, i.e., before diverging the scale of coalescent time unit would be $t = \frac{t'}{2N_e}$ and after diverging it would be $t = \frac{t'}{N_e}$.

Remark 5 In fact, we can simplify $g_{21}(t_i)$ for some coalescent time interval $t_i > 0$ and it can be written as

$$g_{21}(t_i) = 1 - \exp(-t_i).$$

Before we show the probability that any of these three distributions between the caterpillar species tree and gene trees generated from the coalescent process equals to zero, we have to define some notation.

To consider this problem, we need to count the number of cases of $M \in \mathbb{N}$ branches with $N \in \mathbb{N}$ lineages in total. Let $C_{N,M}$ be the number of cases that N lineages coalesce to M lineages. We call the number of lineages in a specific branch the "branch degree". Obviously, the answer depends on if we consider the orders among branches with the same branch degree. If we consider the two figures in Fig. 4 as different cases, then it is not very difficult to obtain that $C_{N,M} = \frac{\prod_{i=2}^{N}\binom{i}{2}}{\prod_{i=2}^{M}\binom{i}{2}}$. However, it will be more complicated if we consider them as the same case. We need to first enumerate all possible ordered M branch degrees (number of lineages coalesce in the branch), then sum up the number of cases for each ordered branch degrees. For example, when $N = 5$ and $M = 3$, we have two possible ordered branch degrees (113) and (122); since for we have $\binom{5}{3} * (2 \cdot 3 - 3)!! = 30$ cases for (113), and $\binom{5}{2}\binom{3}{2}/2 = 15$ cases for (122), we have 45 cases in total.

Define $\mathcal{D}_{M,N} = \{(w_1, w_2, \dots, w_M) \in \mathbb{Z}_+^M : \sum_{l=1}^{M} w_l = N,\ w_1 \leq w_2 \leq \dots \leq w_M\}$ as the set of all possible ordered branch degrees. It is trivial to prove that we can enumerate all elements in $\mathcal{D}_{M,N}$ without duplication in the following way:

$$\mathcal{D}_{M,N} = \left\{ (w_1, w_2, \dots, w_M) \in \mathbb{Z}_+^M : w_1 = 1, 2, \dots, \left\lfloor \frac{N}{M} \right\rfloor,\ w_2 = 1, 2, \dots, \right.$$

$$\left. \left\lfloor \frac{N - w_1}{M-1} \right\rfloor, \dots, w_{M-1} = 1, 2, \dots, \left\lfloor \frac{N - \sum_{l=1}^{M-2} w_l}{M-1} \right\rfloor,\ w_M = N - \sum_{l=1}^{M-1} w_l \right\},$$

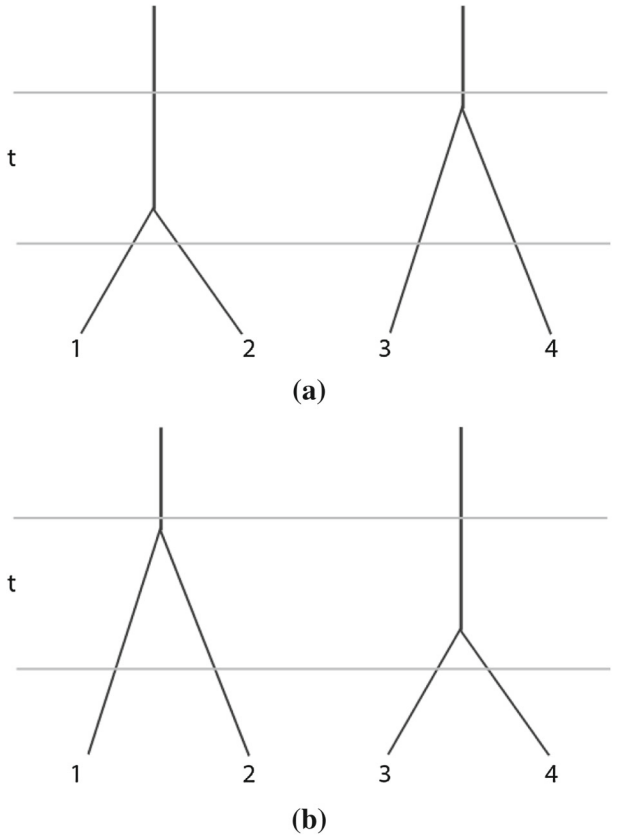

Fig. 4 4 Lineages coalesce to 2 lineages with the same topology 12|34

where "$\lfloor\cdot\rfloor$" gives the largest integer that is smaller than a specific real number. We can define an 1-1 mapping over $\mathcal{D}_{M,N}$ such that $\forall \mathbf{w} = (w_1, w_2, \ldots, w_M) \in \mathcal{D}_{M,N}$, $\mathbf{w}$ maps to two vectors $\mathbf{n}(\mathbf{w}) = (n_0, n_1, \ldots, n_l) \in \mathbb{Z}_+^{l+1}$ and $\mathbf{u}(\mathbf{w}) = (u_0, u_1, \ldots, u_l) \in \mathbb{Z}_+^{l+1}$ which satisfy

$$\mathbf{w} = (\underbrace{n_0, \ldots, n_0}_{u_0 \text{ many}}, \underbrace{n_1, \ldots, n_1}_{u_1 \text{ many}}, \ldots, \underbrace{n_l, \ldots, n_l}_{u_l \text{ many}}).$$

where $n_0 = 1 < n_1 < n_2 < \cdots < n_l$. Notice that this implies $\sum_{\alpha=0}^{l} u_\alpha = M$ and $\sum_{\alpha=0}^{l} u_\alpha n_\alpha = N$.

Lemma 3

$$C_{M,N} = \sum_{\mathbf{n}(\mathbf{w}),\mathbf{u}(\mathbf{w}):\mathbf{w}\in\mathcal{D}_{M,N}} \left\{ \frac{N!}{u_0!} \prod_{\alpha=1}^{l} \frac{((2n_\alpha - 3)!!)^{u_\alpha}}{u_\alpha!(n_\alpha!)^{u_\alpha}} \right\}.$$

Proof Consider $\mathbf{n}(\mathbf{w})$ and $\mathbf{u}(\mathbf{w})$ of an arbitrary $\mathbf{w} \in \mathcal{D}_{M,N}$. We have u_α branches with degree n_α, $\alpha = 0, 1, \ldots, l$. For each branch with degree n_α, we have $(2n_\alpha - 3)!!$ different tree topologies. Notice that we don't consider the permutation among the u_α branches with degree n_α. Thus the number of cases that we choose first u_1 branches with degree n_1 is:

$$\begin{aligned}
&\frac{\binom{N}{n_1}\binom{N-n_1}{n_1}\cdots\binom{N-(u_1-1)n_1}{n_1}[(2n_1-3)!!]^{u_1}}{u_1!} \\
&= \frac{N!}{n_1!\cancel{(N-n_1)!}} \cdot \frac{\cancel{(N-n_1)!}}{n_1!\cancel{(N-2n_1)!}} \cdots \frac{\cancel{(N-(u_1-1)n_1)!}}{n_1!(N-u_1n_1)!} \cdot \frac{[(2n_1-3)!!]^{u_1}}{u_1!} \\
&= \frac{N!}{(n_1!)^{u_1}(N-u_1n_1)!}\frac{[(2n_1-3)!!]^{u_1}}{u_1!}.
\end{aligned}$$

□

Therefore, consider the rest branches, the total number of cases, $C_{M,N}$, is:

$$\begin{aligned}
&\frac{\binom{N}{n_1}\cdots\binom{N-(u_1-1)n_1}{n_1}[(2n_1-3)!!]^{u_1}}{u_1!} \cdot \frac{\binom{N-u_1n_1}{n_2}\cdots\binom{N-u_1n_1-(u_2-1)n_2}{n_2}[(2n_2-3)!!]^{u_2}}{u_2!} \\
&\cdots \frac{\binom{N-\sum_{\alpha=1}^{l-1}u_\alpha n_\alpha}{n_l}\cdots\binom{N-\sum_{\alpha=1}^{l-1}u_\alpha n_\alpha-(u_l-1)n_l}{n_l}[(2n_l-3)!!]^{u_l}}{u_l!} \\
&= \frac{N!}{(n_1!)^{u_1}\cancel{(N-u_1n_1)!}}\frac{[(2n_1-3)!!]^{u_1}}{u_1!} \cdot \frac{\cancel{(N-u_1n_1)!}}{(n_2!)^{u_2}\cancel{(N-\sum_{\alpha=1}^{2}u_\alpha n_\alpha)!}}\frac{[(2n_2-3)!!]^{u_2}}{u_2!} \\
&\cdots \frac{\cancel{(N-\sum_{\alpha=1}^{l-1}u_\alpha n_\alpha)!}}{(n_l!)^{u_l}(N-\sum_{\alpha=1}^{l}u_\alpha n_\alpha)!}\frac{[(2n_l-3)!!]^{u_l}}{u_l!} \\
&= N! \cdot \frac{[(2n_1-3)!!]^{u_1}}{(n_1!)^{u_1}u_1!} \cdot \frac{[(2n_2-3)!!]^{u_2}}{(n_2!)^{u_2}u_2!} \cdots \frac{[(2n_l-3)!!]^{u_l}}{(n_l!)^{u_l}u_l!}\frac{1}{(u_0)!}.
\end{aligned}$$

Example 1 The following table gives the values of $C_{M,N}$ when $N \le 6$:

N / M	1	2	3	4	5	6
1	1	1	3	15	105	945
2		1	3	15	105	945
3			1	6	45	420
4				1	10	105
5					1	15
6						1

Take $N = 6$, $M = 3$ for example. There are 3 possible ordered branch degrees:

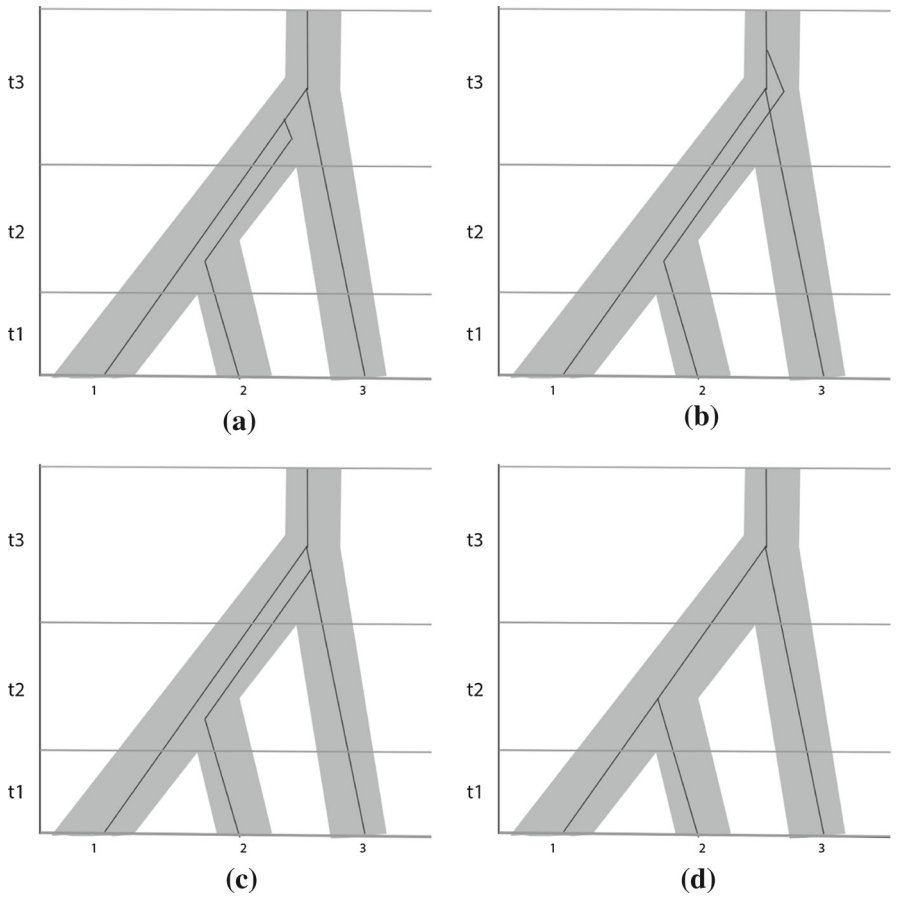

Fig. 5 All possible gene trees for the fixed species tree 12|3

1. $\mathbf{w} = (114), \mathbf{n} = (14), \mathbf{u} = (21)$, number of cases: $\frac{6!}{2!} \cdot \frac{[(2*4-3)!!]^1}{1!(4!)^1} = 225$;
2. $\mathbf{w} = (123), \mathbf{n} = (123), \mathbf{u} = (111)$, number of cases: $\frac{6!}{1!} \cdot \frac{[(2*2-3)!!]^1}{1!(2!)^1} \cdot \frac{[(2*3-3)!!]^1}{1!(3!)^1} = 180$;
3. $\mathbf{w} = (222), \mathbf{n} = (12), \mathbf{u} = (03)$, number of cases: $\frac{6!}{0!} \cdot \frac{[(2*2-3)!!]^3}{3!(2!)^3} = 15$.

So $C_{6,3} = 225 + 180 + 15 = 420$.

For n species, $n - 1$ coalescences should happen during coalescent times $t_1, t_2, \ldots, t_n$. Here, we call the pattern of how these coalescences (regardless of which lineages are K_{th} coalescence) distributed over the coalescent times, i.e. in which coalescent time does the k_{th} coalescent happen, the coalescent timeline. When the gene tree completely matches the species tree, we know that the tree topology of the gene tree is fixed, i.e. the pattern and ordering of coalescence are fixed. This means that the only thing we need to think about is the coalescent timeline. Let's first see a simple example.

Recall $g_{ij}(t)$ is the probability that i lineages coalesce to j lineages in time t.

Example 2 Consider 3 species. Fix the species tree to be 12|3. Figure 5 gives all possible gene trees based on this species tree.

We can compute the probabilities of these trees as following and verify them by summing up to 1:

– Cases for Fig. 5a–c:

$$\begin{aligned}&\Pr((1,2) \text{ in } t_3, (12,3) \text{ in } t_3) = \Pr((1,3) \text{ in } t_3, (13,2) \text{ in } t_3)\\ &= \Pr((2,3) \text{ in } t_3, (23,1) \text{ in } t_3) = \frac{1}{C_{3,1}} g_{22}(t_2) = \frac{1}{3} e^{-t_2}.\end{aligned}$$

Notice that we have $\frac{1}{C_{3,1}}$ here because all these trees share the same coalescent timeline (both coalescences happen in t_3), and we have $C_{3,1}$ cases in t_3 where 3 lineages coalesce to 1 lineage;

– Case for Fig. 5d: $\Pr((1,2) \text{ in } t_2, (12,3) \text{ in } t_3) = g_{21}(t_2) = 1 - e^{-t_2}$.

In this example, $\Pr(d(T_s, T_e) = 0) = \Pr((1,2) \text{ in } t_3, (12,3) \text{ in } t_3) + \Pr((1,2) \text{ in } t_2, (12,3) \text{ in } t_3) = 1 - \frac{2}{3} e^{-t_2}$.

Since for each coalescent timeline, there is only one case giving a gene tree which completely matches the species tree, all we need to do is to enumerate the coalescent timeline and compute the probability for each of them.

Theorem 5 *For n species,*

$$\Pr(d(T_s, T_e) = 0) = \sum_{i_2=0}^{1} \sum_{i_3=i_2}^{2} \cdots \sum_{i_k=i_{k-1}}^{k-1} \cdots \sum_{i_{n-1}=i_{n-2}}^{n-2} \left\{ \left[\prod_{k=2}^{n-1} \frac{g_{k-i_{k-1},k-i_k}(t_k)}{C_{k-i_{k-1},k-i_k}} \right] \cdot \frac{1}{C_{n-i_{n-1},1}} \right\}, \tag{5}$$

where $i_1 = 0$.

Proof Several requirements when we enumerate the coalescent timelines: (1) no coalescent in time t_1; (2) if the i_{th} coalescence happens in time t_{k_i}, then $i+1 \le k_i \ge n$; (3) if the i_{th} and j_{th} coalescences happen in time t_{k_i} and t_{k_j} respectively and $i < j$, then $k_i \le k_j$ (otherwise the gene tree will have a different tree topology with the species tree); (4) all lineages coalescent to one in time t_n.

In Eq. 5, every choice of $(i_1, i_2, \ldots, i_{n-1})$ gives a possible coalescent timeline: i_k coalescences happen before or during time t_k, $k = 1, 2, \ldots, n-1$, and $(n - i_{n-1})$ coalescences happen during time t_n. It is trivial to see that these choices enumerate all possible coalescent timelines without duplicates.

Now consider a specific $(i_1, i_2, \ldots, i_{n-1})$. Then during time $t_k, k = 2, 3, \ldots, n-1$, since the input has k species with i_{k-1} coalescences, i.e. $k - i_{k-1}$ lineages, and the output has k species with i_k coalescences, i.e. $k - i_k$ lineages, the probability that the gene tree completely agrees with the species tree is $\frac{g_{k-i_{k-1},k-i_k}(t_k)}{C_{k-i_{k-1},k-i_k}}$ (see example in Fig.

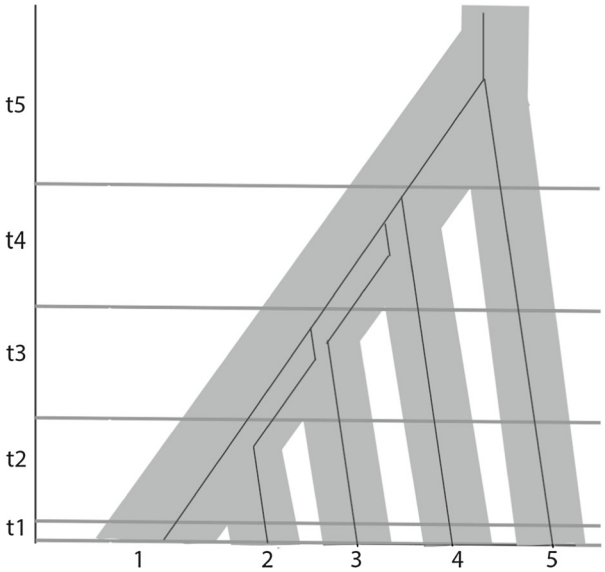

Fig. 6 5 species with timeline: $(i_1, i_2, i_3, i_4) = (0, 0, 1, 3)$. $i_2 - i_1 = 0$ coalescent happened in t_2; $i_3 - i_2 = 1$ coalescent happened in t_3; $i_4 - i_3 = 2$ coalescents happened in t_4. In time t_3, we have $3 - i_2 = 3$ lineages coming and $3 - i_3 = 2$ lineages coming out, so the probability that we get exactly the same topology as this figure during time t_3 is $\frac{g_{32}(t_3)}{C_{3,2}}$

6). During time t_n, we left $n - i_{n-1}$ lineages and they should coalesce to one, so the probability should be $\frac{1}{C_{n-i_{n-1},1}}$.

□

Example 3 There are five cases for $n = 4$ so that gene tree completely matches the species tree. We apply Theorem 5 for $n = 4$ in and obtain the following probabilities for each of the cases:

1. Coalescents $(1, 2)$ in t_4; $(12, 3)$ in t_4; $(123, 4)$ in t_4 (see Fig. 7a). Probability is $\frac{1}{15} g_{22}(t_2) g_{33}(t_3)$;
2. Coalescents $(1, 2)$ in t_3; $(12, 3)$ in t_4; $(123, 4)$ in t_4 (see Fig. 7b). Probability is $\frac{1}{9} g_{22}(t_2) g_{32}(t_3)$;
3. Coalescents $(1, 2)$ in t_3; $(12, 3)$ in t_3; $(123, 4)$ in t_4 (see Fig. 7c). Probability is $\frac{1}{3} g_{22}(t_2) g_{31}(t_3)$;
4. Coalescents $(1, 2)$ in t_2; $(12, 3)$ in t_4; $(123, 4)$ in t_4 (see Fig. 7d). Probability is $\frac{1}{3} g_{21}(t_2) g_{22}(t_3)$;
5. Coalescents $(1, 2)$ in t_2; $(12, 3)$ in t_3; $(123, 4)$ in t_4 (see Fig. 7e). Probability is $g_{21}(t_2) g_{21}(t_3)$;

Then we have formula:

$$\Pr(d(T_s, T_e) = 0) = \frac{1}{15} g_{22}(t_2) g_{33}(t_3) + \frac{1}{9} g_{22}(t_2) g_{32}(t_3) + \frac{1}{3} g_{22}(t_2) g_{31}(t_3) + \frac{1}{3} g_{21}(t_2) g_{22}(t_3) + g_{21}(t_2) g_{21}(t_3).$$

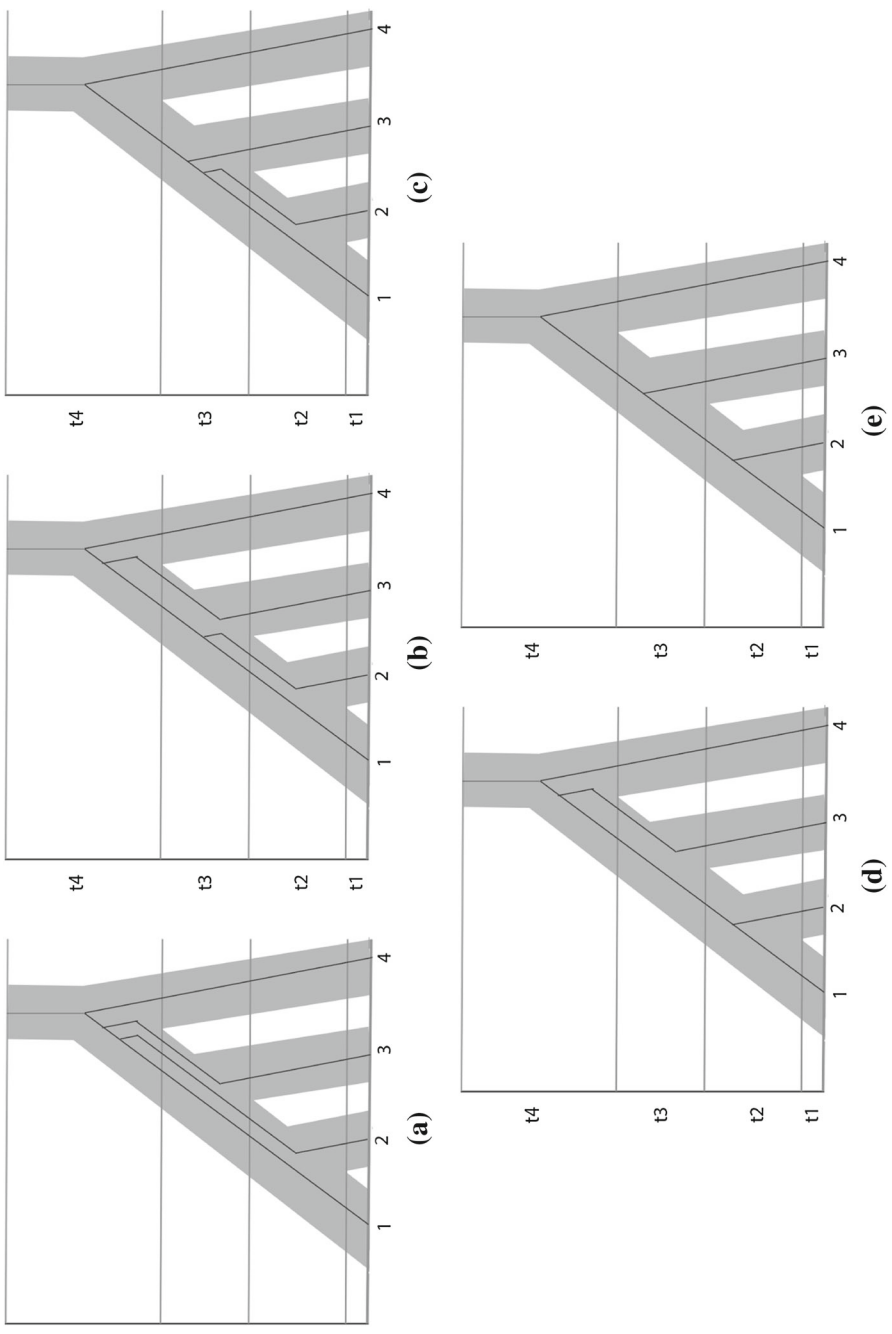

Fig. 7 Gene trees with $d(T_s, T_e) = 0$ and their coalescent timelines (i_1, i_2, i_3)

By Theorem 5, if we have larger t_k for $k = 1, \ldots, n$, then we have higher probability that the species tree T_s and its gene tree T_g generated under the coalescent given T_s have the same tree topology. In addition, since k-IC is the l_∞ norm of the vector in $\mathbb{R}^{\binom{n}{2}}$, the path difference is the l_2 norm of the vector in $\mathbb{R}^{\binom{n}{2}}$, and the edge difference is the l_1 norm of the vector in $\mathbb{R}^{\binom{n}{2}}$, k-IC distance tree metric can be used for the upper bound for the path difference tree metric and the edge difference tree metric by Remark 3. Thus, by Lemmas 1 and 2, if we have larger t_k for $k = 1, \ldots, n$, then the distributions of tree distance metric d_e, d_p and d_k between T_s and T_g are skewed from right.

5 Simulations

First we have conducted simulations study on the three tree distances, the edge difference, path difference, and precise K-IC distances between two unrooted random trees with 12 leaves. We have conducted a simulation study similar to what (Steel and Penny 1993) did (Fig. 6 on their paper). We generated 10,000 unrooted random trees with 12 leaves using the function `rtree` from `R` package `ape` (Paradis et al. 2004). Then for each distance measure d_e, d_p, d_k we computed a histogram. In order to compare a histogram with each other we normalized the distances so that they scale from 0 to 10. The results are shown in Fig. 8. We also conducted the same simulations with the function `rcoal` from `ape` and we have obtained basically the same results.

In the second simulation part, we conducted a simulation study on the distributions of d_e, d_p, d_k between the caterpillar species tree and a random gene tree generated from the coalescent process with the species tree. We used the software `Mesquite` (Maddison and Maddison 2011) to generate caterpillar species trees with 5 leaves, 6 leaves, 7 leaves and 8 leaves, respectively under the Yule process. Then we simulated 10,000 gene trees within each species tree. For all the trees in the simulation, they have the same parameters, that is the effective population size $N_e = 30{,}000$ and species depth = 1,000. For each kind of trees with certain number of leaves, we then calculated three different kinds of distances between the gene trees and species trees. Table 1 shows the proportions of 0 and 1 distances in each of the three distances for the rooted trees with 5 leaves, 6 leaves, 7 leaves and 8 leaves, respectively. Figures 9, 10, and 11 show the histograms of three kinds of distances for trees with 5 leaves, 6 leaves, 7 leaves and 8 leaves, respectively.

6 Discussion

While many tree distances measures between trees are hard to compute (see Remark 1) tree distances d_e, d_p, d_k can be computed in polynomial time in n. Today, we can generate huge numbers of DNA sequences from genomes using new generation sequencing techniques and they can generate tens of millions base pairs of DNA sequences. In order to conduct phylogenomics analysis on genome data sets we need fast tree distances, such as d_e, d_p, d_k. However, in order to understand statistical

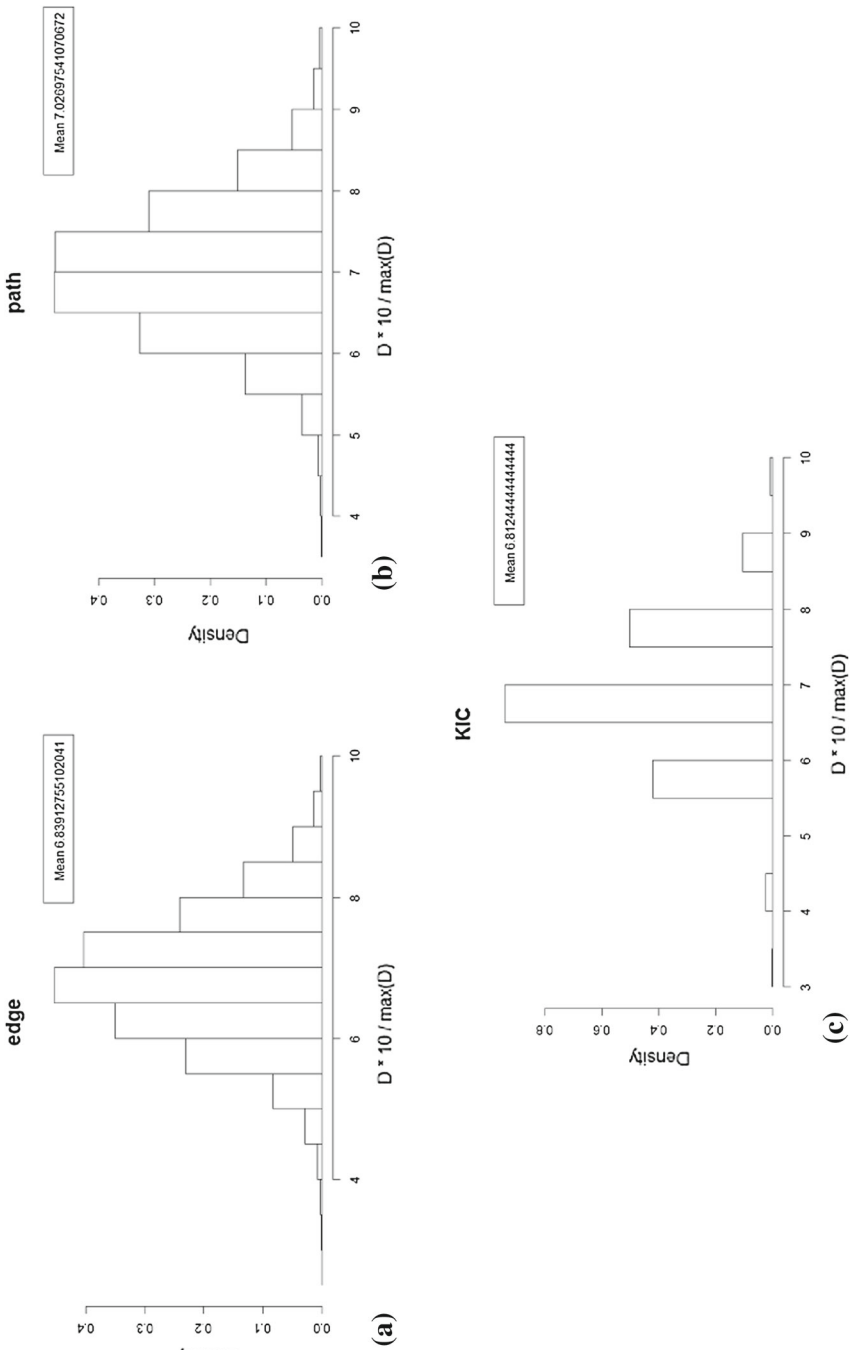

Fig. 8 We generated 10,000 random trees using the function rtree from ape

Table 1 The proportions of 0 and 1 distances in each of the three distances d_e, d_p, d_k for the rooted trees with 5 leaves, 6 leaves, 7 leaves and 8 leaves

	Sample proportion	Mean distance	Standard deviation
5 leaves			
$d_k = 0$	0.9543	0.0457	0.2088
$d_k = 1$	0.0457		
$d_e = 0$	0.9543	0.2742	1.2531
$d_e = 1$	0		
$d_p = 0$	0.9543	0.1119	0.5116
$d_p = 1$	0		
6 leaves			
$d_k = 0$	0.9007	0.1025	0.3137
$d_k = 1$	0.0961		
$d_e = 0$	0.9007	0.8200	2.4899
$d_e = 1$	0		
$d_p = 0$	0.9007	0.2869	0.8682
$d_p = 1$	0		
7 leaves			
$d_k = 0$	0.4824	0.6842	0.7420
$d_k = 1$	0.3516		
$d_e = 0$	0.4824	6.5920	6.7687
$d_e = 1$	0		
$d_p = 0$	0.4824	1.9531	1.9685
$d_p = 1$	0		
8 leaves			
$d_k = 0$	0.0760	1.8490	0.9002
$d_k = 1$	0.2639		
$d_e = 0$	0.0760	20.2859	9.0730
$d_e = 1$	0		
$d_p = 0$	0.0760	5.1175	2.0716
$d_p = 1$	0		

phylogenomics analysis on genome data sets with these tree distances, we have to understand distribution of these distances.

In this paper we have shown some theoretical and simulation results on the distributions of tree distances d_e, d_p, d_k between unrooted random trees with n leaves and between the caterpillar species tree and a random rooted gene tree with n leaves generated from the coalescent process with the species tree.

The distributions of tree distances d_e, d_p, d_k between unrooted random trees with n leaves seem to be symmetric and we have conducted some goodness of fit test with the Gaussian distribution. However, the null hypothesis (the distribution fits with the Gaussian distribution) seems to be rejected (with the number of trees equals to 10,000),

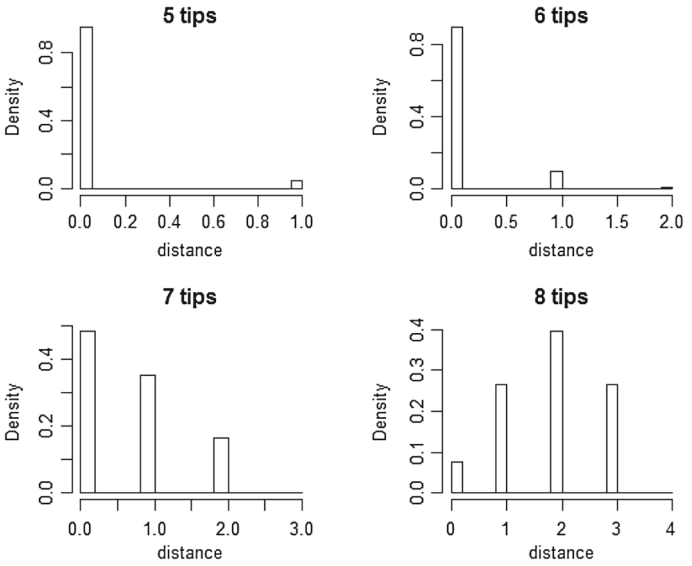

Fig. 9 Histogram of d_k for the caterpillar species tree and a random tree generated from the coalescent process

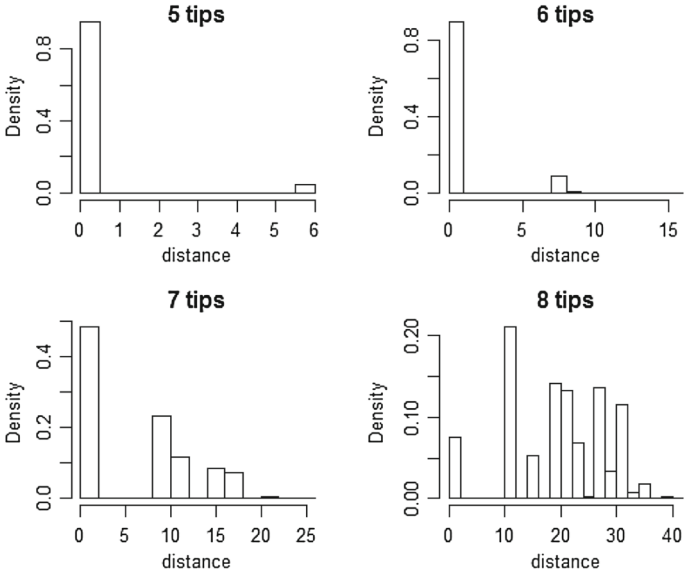

Fig. 10 Histogram of d_e for the caterpillar species tree and a random tree generated from the coalescent process

so it would be interesting and useful to know the asymptotic distributions of d_e, d_p, d_k between unrooted random trees with n leaves.

In Theorem 5, we have shown explicitly the probability of the tree distance d_e, d_p, d_k between caterpillar species tree with n leaves and a random gene tree

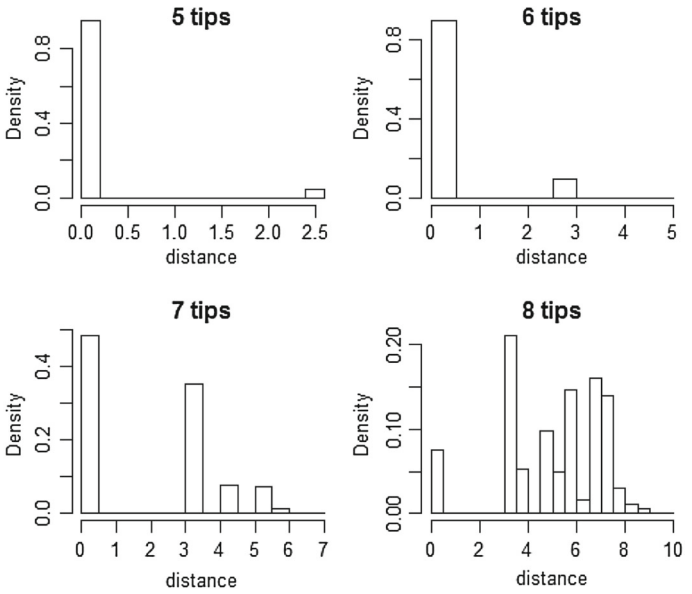

Fig. 11 Histogram of d_p for the caterpillar species tree and a random tree generated from the coalescent process

with n leaves distributed with the coalescent process with the species tree equals to zero. Note here the species tree is assumed to be caterpillar because d_k between two trees can reach its upper bound only if one of them is caterpillar. Figures 9, 10 and 11 show us that when the sizes of trees get larger, the centers and variation of non-zero distances also become larger, but zero is the only distance value that always guarantee a positive probability for all three types of distances. We are also interested in computing the probability of d_k being one, which is generally zero for d_e and d_p (see Table 1). However we do not know many aspects of the tree distance d (one of the distances d_e, d_p, d_k) between them as $n \to \infty$. Thus, we have the following questions.

Problem 1 Consider the tree distances d_e, d_p, d_k between caterpillar species tree with n leaves and a random gene tree with n leaves distributed with the coalescent process with the species tree. What is the expectation of the tree distance d (one of the distances d_e, d_p, d_k) between them? How about variance? Can we say anything about the expectation asymptotically?

Acknowledgements The authors would like to thank the referees for very useful comments to improve the manuscript.

References

Allen, B., Steel, M. (2001). Subtree transfer operations and their induced metrics on evolutionary trees. *Annals of Combinatorics*, *5*(1), 1–15.

Arnaoudova, E., Haws, D., Huggins, P., Jaromczyk, J. W., Moore, N., Schardl, C., et al. (2010). Statistical phylogenetic tree analysis using differences of means. *Frontier Psychiatry*, *1*(47).

Betancur, R., Li, C., Munroe, T., Ballesteros, J., Ortí, G. (2013). Addressing gene tree discordance and non-stationarity to resolve a multi-locus phylogeny of the flatfishes (teleostei: Pleuronectiformes). *Systematic Biology*,. doi:10.1093/sysbio/syt039.

Bollback, J., Huelsenbeck, J. (2009). Parallel genetic evolution within and between bacteriophage species of varying degrees of divergence. *Genetics*, *181*(1), 225–234.

Brito, P., Edwards, S. (2009). Multilocus phylogeography and phylogenetics using sequence-based markers. *Genetica*, *135*, 439–455.

Brodal, G., Fagerberg, R., Pedersen, C. N. (2001). Computing the quartet distance between evolutionary trees in time nlog2n. *Algorithmica*, 731–742.

Carling, M., Brumfield, R. (2008). Integrating phylogenetic and population genetic analyses of multiple loci to test species divergence hypotheses in passerina buntings. *Genetics*, *178*, 363–377.

Carstens, B. C., Knowles, L. L. (2007). Estimating species phylogeny from gene-tree probabilities despite incomplete lineage sorting: an example from melanoplus grasshoppers. *Systematic Biology*, *56*, 400–411.

Coons, J. Rusinko, J. (2014). Combinatorics of k-interval cospeciation for cophylogeny. http://arxiv.org/pdf/1407.6605.pdf (preprint)

Dasgupta, B., He, X., Jiang, T., Li, M., Tromp, J., Zhang, L. (1997). On computing the nearest neighbor interchange distance. In Proceedings of DIMACS Workshop on Discrete Problems with Medical Applications (pp. 125–143) (press).

Degnan, J., Salter, L. (2005a). Gene tree distribtutions under the coalescent process. *Evolution*, *59*(1), 24–37.

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

Edwards, S. (2009). Is a new and general theory of molecular systematics emerging? *Evolution*, *63*, 1–19.

Edwards, S., Liu, L., Pearl, D. (2007). High-resolution species trees without concatenation. *Proceedings of the National Academy of Sciences USA*, *104*, 5936–5941.

Graham, M., Kennedy, J. (2010). A survey of multiple tree visualisation. *Information Visualization*, *9*, 235–252.

Heled, J., Drummond, A. (2011). Bayesian inference of species trees from multilocus data. *Molecular Biology and Evolution*, *27*(3), 570–580.

Hickey, G., Dehne, F., Rau-Chaplin, A., Blouin, C. (2008). SPR distance computation for unrooted trees. *Evolutionary Bioinformatics Online*, *4*, 17–27.

Hillis, D. M., Heath, T. A., St. John, K. (2005). Analysis and visualization of tree space. *Systematic Biology*, *54*(3), 471–482.

Holmes, S. (2007). *Statistical Approach to Tests Involving Phylogenies*. New York: Oxford University Press.

Huggins, P., Owen, M., Yoshida, R. (2012). First steps toward the geometry of cophylogeny. In The Proceedings of the Second CREST-SBM International Conference "Harmony of Gröbner Bases and the Modern Industrial Society" (pp. 99–116).

Maddison, W. P. (1997). Gene trees in species trees. *Systematic Biology*, *46*(3), 523–536.

Maddison, W. P., Knowles, L. L. (2006). Inferring phylogeny despite incomplete lineage sorting. *Systematic Biology*, *55*, 21–30.

Maddison, W. P. Maddison, D. R. (2011). Mesquite: a modular system for evolutionary analysis. version 2.75.

Mossel, E., Roch, S. (2010). Incomplete lineage sorting: consistent phylogeny estimation from multiple loci. *IEEE/ACM Transactions on Computational Biology and Bioinformatics*, *7*(1), 166–171.

Pamilo, P., Nei, M. (1988). Relationships between gene trees and species trees. *Molecular Biology and Evolution*, *5*, 568–583.

Paradis, E., Claude, J., Strimmer, K. (2004). APE: analyses of phylogenetics and evolution in R language. *Bioinformatics*, *20*, 289–290.

Robinson, D. F., Foulds, L. R. (1981). Comparison of phylogenetic trees. *Mathematical Biosciences*, *53*, 131–147.

Rosenberg, N. (2002). The probability of topological concordance of gene trees and species trees. *Theoretical Population Biology*, *61*, 225–247.

Rosenberg, N. A. (2003). The shapes of neutral gene genealogies in two species: probabilities of monophyly, paraphyly, and polyphyly in a coalescent model. *Evolution*, *57*, 1465–1477.

RoyChoudhury, A., Felsenstein, J., Thompson, E. A. (2008). A two-stage pruning algorithm for likelihood computation for a population tree. *Genetics*, *180*, 1095–1105.

Semple, C. Steel, M. (2003). Phylogenetics, vol. 24 of Oxford Lecture Series in mathematics and its applications. Oxford: Oxford University Press.

Steel, M., Penny, D. (1993). Distributions of tree comparison metrics-some new results. *Systematic Biology*, *42*(2), 126–141.

Takahata, N. (1989). Gene genealogy in 3 related populations: consistency probability between gene and population trees. *Genetics*, *122*, 957–966.

Takahata, N., Nei, M. (1990). Allelic genealogy under overdominant and frequency-dependent selection and polymorphism of major histocompatibility complex loci. *Genetics*, *124*, 967–978.

Tavaré, S. (1984). Line-of-descent and genealogical processes, and their applications in population genetics models. *Theoretical Population Biology*, *26*, 119–164.

Thompson, K., Kubatko, L. (2013). Using ancestral information to detect and localize quantitative trait loci in genome-wide association studies. *BMC Bioinformatics*, *14*, 200.

Weyenberg, G., Huggins, P., Schardl, C., Howe, D., Yoshida, R. (2014). kdetrees: non-parametric estimation of phylogenetic tree distributions. *Bioinformatics*, *30*(16), 2280–2287.

Williams, W. T., Clifford, H. T. (1971). On the comparison of two classifications of the same set of elements. *Taxon*, *20*, 519–522.

Yu, Y., Warnow, T., Nakhleh, L. (2011). Algorithms for mdc-based multi-locus phylogeny inference: Beyond rooted binary gene trees on single alleles. *Journal of Computational Biology*, *18*(11), 1543–1559.