JOURNAL OF COMPUTATIONAL BIOLOGY
Volume 18, Number 1, 2011

Pp. 1–15
DOI: 10.1089/cmb.2010.0102

Research Articles

Consistency Properties of Species Tree Inference by Minimizing Deep Coalescences

CUONG V. THAN[1] and NOAH A. ROSENBERG[1,2]

ABSTRACT

Methods for inferring species trees from sets of gene trees need to account for the possibility of discordance among the gene trees. Assuming that discordance is caused by incomplete lineage sorting, species tree estimates can be obtained by finding those species trees that minimize the number of "deep" coalescence events required for a given collection of gene trees. Efficient algorithms now exist for applying the minimizing-deep-coalescence (MDC) criterion, and simulation experiments have demonstrated its promising performance. However, it has also been noted from simulation results that the MDC criterion is not always guaranteed to infer the correct species tree estimate. In this article, we investigate the consistency of the MDC criterion. Using the multipscies coalescent model, we show that there are indeed anomaly zones for the MDC criterion for asymmetric four-taxon species tree topologies, and for all species tree topologies with five or more taxa.

Key words: algorithms, combinatorial optimization, combinatorics, computational molecular biology, phylogenetic trees.

1. INTRODUCTION

IT IS WELL KNOWN THAT, FOR A VARIETY OF REASONS such as horizontal gene transfer, gene duplication and loss, and incomplete lineage sorting, gene trees can differ from each other and from the species tree along whose branches they have evolved (Degnan and Rosenberg, 2009; Maddison, 1997; Nichols, 2001). Consequently, methods for inferring species trees from sets of gene trees need to consider gene tree discordance in order to obtain reliable estimates.

Approaches to resolving the species tree/gene tree discordance problem in phylogenetic inference can be classified as either nonparametric (e.g., democratic vote, consensus, and parsimony-based) or parametric (e.g., likelihood and Bayesian). In general, nonparametric methods are faster than parametric methods, and hence they are computationally preferable for analyzing large datasets. One of the main concerns about these methods, however, is their potential for inconsistency. Under a specific model for the evolution of gene trees along the branches of species trees, a method is consistent if for each collection of values of the model parameters—the species tree topology and its branch lengths—the method produces a correct estimate of the species tree in the limit as the number of sampled gene trees goes to infinity. Recently, inconsistency results have been reported for several nonparametric methods,

[1]Center for Computational Medicine and Bioinformatics and [2]Department of Human Genetics and the Life Sciences Institute, University of Michigan, Ann Arbor, Michigan.

including democratic vote and several consensus methods. For example, Degnan and Rosenberg (2006) have shown that for asymmeric species trees with four leaves and for any species tree with at least five leaves, there exist species tree branch lengths such that the most likely gene tree topology under the multispecies coalescent model (Degnan and Rosenberg, 2009), the ''democratic vote'' topology, differs from the species tree topology. The greedy consensus method, which reconstructs the species tree by sequentially adding the most frequent clade compatible with all previously included clades, and which is not based specifically on coalescent principles, has also been proven to be inconsistent (Degnan et al., 2009). In contrast, several methods for inferring species trees from gene trees that make use of elements of the coalescent model, such as STAR (Liu et al., 2009), STEAC (Liu et al., 2009), and GLASS (Liu et al., 2010; Mossel and Roch, 2010), have been shown to be consistent under the multispecies coalescent model.

Maddison (1997) introduced a parsimony criterion for inferring species trees from gene trees by minimizing deep coalescences (MDC), and several exact algorithms and heuristics for implementing this criterion have recently been developed (Bansal et al., 2010; Than and Nakhleh, 2009). Unlike the democratic vote or greedy consensus methods, which provide algorithms for inferring species trees from collections of gene trees without taking into account the process by which the gene trees have been produced, the MDC criterion relies on an understanding of the specific nature of the way in which incomplete lineage sorting occurs. Thus, it is a natural candidate for species tree inference when discordance among gene trees is caused by incomplete lineage sorting. Simulation studies have suggested a high degree of accuracy of species tree estimates obtained by this criterion (Maddison and Knowles, 2006; Than and Nakhleh, 2009).

As noted by Than and Nakhleh (2009), however, it has been observed that in some cases, the MDC criterion does not reconstruct the correct species tree. As no theoretical results concerning consistency properties of the criterion have yet been reported, in this paper we investigate whether it is consistent under the multispecies coalescent model. We show that if gene lineages have evolved according to the multispecies coalescent model, then the MDC criterion is inconsistent. In other words, for certain combinations of species tree topologies and branch lengths, the MDC criterion infers an incorrect species tree topology in the limit as the number of sampled genes increases without bound.

2. THE MINIMIZING-DEEP-COALESCENCE CRITERION

Although a variety of reasons can explain why gene trees can disagree with the species tree that contains them, we assume throughout this article that incomplete lineage sorting, or deep coalescence, is the only source for the discordance. Looking backward in time, the discordance between a gene tree and a species tree occurs because gene lineages can persist deeper than speciation events, providing opportunities for them to coalesce in an order different from the order of speciation events.

2.1. The deep coalescence cost

To measure the severity of the topological disagreement between a gene tree and a species tree, we use the deep coalescence cost, first introduced by Maddison (1997). Given a binary, rooted gene tree T and species tree S on a taxon set X, the deep coalescence cost for reconciling T within S is computed as follows. Each node v of T is mapped to its most recent common ancestor (MRCA) node in S, that is, the most recent node in S whose descendant leaf set (in S) contains all of the descendant leaves of v in T (Fig. 1). For each internal branch e of S, let $\mathrm{xl}_S(T, e)$ be the number of gene lineages at the ''top'' of branch e minus 1; $\mathrm{xl}_S(T, e)$ is also called the number of ''extra'' lineages in e. The deep coalescence cost for reconciling T within S is defined as

$$\alpha(T, S) = \sum_{e \in \mathring{E}(S)} \mathrm{xl}_S(T, e), \tag{2.1}$$

where $\mathring{E}(S)$ is the set of internal branches in S.

It is possible to compute the number of extra lineages in an internal branch e of S without using an MRCA mapping between the nodes of T and the nodes of S (Theorem 2 of Than and Nakhleh, 2009). For an internal branch e of the species tree S, let $C_S(e)$ be the label set of the leaves under e (i.e., $C_S(e)$ is the set of leaf labels for the cluster induced by e). A subtree t of T whose leaf set is contained in $C_S(e)$ is

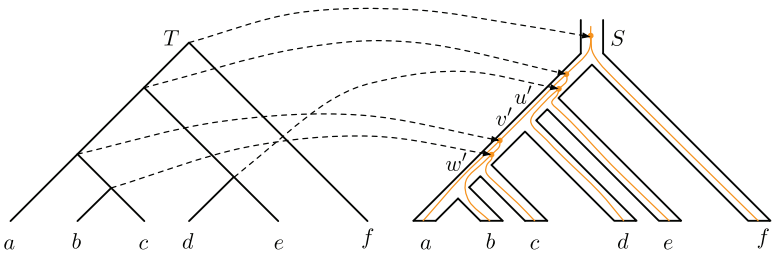

FIG. 1. Computing the deep coalescence cost. Gene tree T is fitted onto species tree S according to a most recent common ancestor (MRCA) mapping. In the figure, mappings between leaves are omitted, and for clearer illustration of how T is reconciled within S, the MRCAs of the internal nodes of T are placed along the branches of S rather than at internal nodes of S. The labels u', v', and w' refer to specific nodes of S. In this example, the minimizing-deep-coalescence (MDC) cost for T and S, $\alpha(T, S)$, is two, the total number of extra lineages in all the branches of S.

maximal with respect to e if it is not a proper subtree of another subtree t' of T whose leaf set is also contained in $C_S(e)$. If k is the number of maximal subtrees of T with respect to e, then the number of extra lineages in e is

$$\mathrm{xl}_S(T, e) = k - 1. \tag{2.2}$$

For example, Figure 1 illustrates that there is one extra lineage in the branch (u', v') of the species tree topology S. We can also obtain this result by noting that (u', v') induces the cluster of leaf labels $C = \{a, b, c, d\}$. There are two subtrees of T whose leaf sets are subsets of C, and that are maximal with respect to the branch (u', v'): $t_1 = (a, (b, c))$ and $t_2 = (d)$. Consequently, from Eq. (2.2) we obtain that the number of extra lineages in (u', v') is 1.

Suppose that we are given a collection G of binary, rooted gene trees on label set X. Denoting by $R(X)$ the set of all possible binary, rooted trees on X, then for each candidate species tree S' in $R(X)$, we compute the deep coalescence cost for reconciling all gene trees in G within S' by evaluating the sum

$$\alpha(G, S') = \sum_{T \in G} \alpha(T, S'), \tag{2.3}$$

where $\alpha(T, S')$ is calculated using Eq. (2.1). Under the MDC criterion, a tree in $R(X)$ whose deep coalescence cost, defined by Eq. (2.3), is the smallest among those of all trees in $R(X)$ is taken as an estimate of the true species tree S. Note that more than one tree can be tied with the smallest deep coalescence cost, and in this case, the MDC criterion randomly chooses one of them as an estimate of S. Efficient algorithms exist for identifying optimal trees under the MDC criterion for a collection of gene trees (Bansal et al., 2010; Than and Nakhleh, 2009).

2.2. *An observation about collections of deep coalescence costs*

Given two trees in $R(X)$ that have the same unlabeled topology, S'_1 and S'_2, let us compare $\{\alpha(T, S'_1) \mid T \in R(X)\}$ and $\{\alpha(T, S'_2) \mid T \in R(X)\}$, the collections of deep coalescence costs for reconciling all trees in $R(X)$ within S'_1 and S'_2, respectively. Because S'_1 and S'_2 have the same unlabeled topology, there exists a permutation π of the taxon set X such that the leaves of S'_1 can be relabeled according to π to obtain S'_2. Denoting by $\pi(T)$, for $T \in R(X)$, the tree obtained from T by applying π to its leaves (so that by our choice of π, $\pi(S'_1) = S'_2$), we have the following facts:

1. If the leaves of both T and S'_1 are relabeled using π, then the MRCA mapping between the nodes of T and S'_1 remains unchanged, and hence $\alpha(T, S'_1) = \alpha(\pi(T), \pi(S'_1)) = \alpha(\pi(T), S'_2)$.
2. Because π is a permutation of X, $\pi(T_1) \neq \pi(T_2)$ if $T_1 \neq T_2$. Moreover, because $R(X)$ is the set of all rooted, binary trees on X, $\{\pi(T) \mid T \in R(X)\} = R(X)$.

These facts imply that $\{\alpha(T, S'_1) \mid T \in R(X)\}$ is equal to $\{\alpha(\pi(T), S'_2) \mid T \in R(X)\}$, which is in turn equal to $\{\alpha(T, S'_2) \mid T \in R(X)\}$.

This observation can be further refined. Note that a tree in $R(X)$ can never be transformed into another tree of different unlabeled topology simply by relabeling its leaves according to a permutation π of X. Therefore, if $R(X)$ is partitioned into subsets $R_1, R_2, \ldots$ of trees having the same unlabeled topology, then for each $i = 1, 2, \ldots, \{\pi(T) \mid T \in R_i\} = R_i$. Thus, $\{\alpha(T, S'_1) \mid T \in R_i\} = \{\alpha(T, S'_2) \mid T \in R_i\}$ for any two tree topologies S'_1 and S'_2 that have the same unlabeled topology. This refined observation, that the

collection of deep coalescence costs of all gene trees having a given unlabeled topology is dependent only on the species tree's *unlabeled* topology, is used in the next section in the proof of the inconsistency of the MDC criterion.

3. INCONSISTENCY OF THE MDC CRITERION

Let S be a binary, rooted species tree on a taxon label set X, and let λ be the vector of the lengths of branches of S. The branch lengths are positive, and are measured in coalescent time units. We assume that gene lineages have evolved along the branches of S following the multispecies coalescent model (Degnan and Rosenberg, 2009). We further assume that one gene lineage is sampled in each species, so that a gene tree and species tree have the same number of lineages, and that gene trees are independent and known with certainty. Under the multispecies coalescent model, the probability of observing a gene tree $T \in R(X)$ given the species tree S, $\Pr(T \mid S, \lambda)$, can be computed using a formula of Degnan and Salter (2005).

For a collection G of binary, rooted gene trees on X, the MDC criterion chooses as an estimate of the species tree S a tree whose deep coalescence cost, defined by Eq. (2.3), is the smallest among those of all trees in $R(X)$. Because the number of gene trees in G is fixed for a given collection G, it is equivalent for the MDC criterion to choose among all trees S' in $R(X)$ a tree with the smallest *mean* deep coalescence cost, defined as $\alpha(G, S')/|G|$. By the strong law of large numbers, as the number of sampled gene trees in G goes to infinity, the mean $\alpha(G, S')/|G|$ approaches with probability 1 the expected value

$$\overline{\alpha}_{S,\lambda}(S') = \sum_{T \in R(X)} \Pr(T \mid S, \lambda)\alpha(T, S'). \tag{3.1}$$

Therefore, in the limit where $|G|$ goes to infinity, a species tree candidate S^* with the smallest expected value $\overline{\alpha}_{S,\lambda}(S^*)$ is chosen as an estimate of the species tree S. We call this tree the asymptotic MDC tree, following the terminology in Degnan et al. (2009). If there is only one asymptotic MDC tree S^*, and S^* differs from S, then the MDC criterion produces an incorrect estimate of S as the number of gene trees increases without bound; that is, the MDC criterion is not statistically consistent. If there is more than one asymptotic MDC tree, we also say that the MDC criterion is not statistically consistent, because in this case it simply randomly picks one of these trees as an estimate of S.

3.1. Trees with three leaves

We first consider trees that have only three leaves. There are three possible labeled rooted, binary trees with three leaves a, b, and c: $S_1 = T_1 = ((a, b), c)$, $S_2 = T_2 = ((a, c), b)$, and $S_3 = T_3 = ((b, c), a)$. Here, for convenience, we refer to these trees as T_1, T_2, and T_3 when using them as gene trees, and as S_1, S_2, and S_3 when using them as species trees. These trees differ only in a permutation of the leaf labels. Therefore, to study the consistency of the MDC criterion, it is sufficient to consider the case in which the true species tree topology is S_1. That is, we can assume the species tree (with branch lengths) is $(S_1, \lambda) = ((a, b): x, c)$, where x is the positive length in coalescent time units of the only internal branch of S_1.

The probabilities of observing the gene trees T_1, T_2, and T_3 are $\Pr(T_1 \mid S_1, \lambda) = 1 - 2e^{-x}/3$ and $\Pr(T_2 \mid S_1, \lambda) = \Pr(T_3 \mid S_1, \lambda) = e^{-x}/3$, respectively (Hudson, 1983; Pamilo and Nei, 1988; Tajima, 1983). It is easy to check that $\alpha(T_1, S_1) = 0$ and $\alpha(T_2, S_1) = \alpha(T_3, S_1) = 1$. Using Eq. (3.1), we have $\overline{\alpha}_{S_1,\lambda}(S_1) = 2e^{-x}/3$. Similarly, we have $\overline{\alpha}_{S_1,\lambda}(S_2) = \overline{\alpha}_{S_1,\lambda}(S_3) = 1 - e^{-x}/3$. Clearly, for positive x, $2e^{-x}/3 < 1 - e^{-x}/3$, implying that S_1 is the only asymptotic MDC tree. Hence, the MDC criterion is statistically consistent for trees with three leaves.

3.2. Trees with four leaves

There are 15 labeled rooted, binary trees on four leaves. This collection of trees can be divided into a set of symmetric trees, $R_1 = \{T_1, \ldots, T_3\}$, and a set of asymmetric trees, $R_2 = \{T_4, \ldots, T_{15}\}$ (Table 1). For convenience, we refer to the ith tree, $i = 1, \ldots, 15$, as T_i when using it as a gene tree, and as S_i when using it as a species tree. Similarly to the case of trees with three leaves, it is sufficient for us to consider only one labeling for each unlabeled species tree topology. We assume that the species tree is either $(S_1, \lambda) = ((a, b):$

TABLE 1. PROBABILITIES AND DEEP COALESCENCE COSTS FOR RECONCILING EACH OF THE 15 ROOTED, BINARY GENE TREES WITH LEAF LABELS *A*, *B*, *C*, AND *D*, GIVEN EITHER THE SPECIES TREE $(S_1, \lambda) = ((a, b): y, (c, d): x)$ OR $(S_4, \lambda) = (((a, b): y, c): x, d)$

Gene tree T_i	$Pr(T_i \mid S_1, \lambda)$	$\alpha(T_i, S_1)$	$Pr(T_i \mid S_4, \lambda)$	$\alpha(T_i, S_4)$
$T_1 = ((a, b), (c, d))$	$1 - \frac{2}{3}(e^{-x} + e^{-y}) + \frac{4}{9}e^{-(x+y)}$	0	$\frac{1}{3}e^{-x} - \frac{1}{6}e^{-(x+y)} - \frac{1}{18}e^{-(3x+y)}$	1
$T_2 = ((a, c), (b, d))$	$\frac{1}{9}e^{-(x+y)}$	2	$\frac{1}{6}e^{-(x+y)} - \frac{1}{18}e^{-(3x+y)}$	2
$T_3 = ((a, d), (b, c))$	$\frac{1}{9}e^{-(x+y)}$	2	$\frac{1}{6}e^{-(x+y)} - \frac{1}{18}e^{-(3x+y)}$	2
$T_4 = (((a, b), c), d)$	$\frac{1}{3}e^{-x} - \frac{5}{18}e^{-(x+y)}$	1	$1 - \frac{2}{3}(e^{-x} + e^{-y}) + \frac{1}{3}e^{-(x+y)} + \frac{1}{18}e^{-(3x+y)}$	0
$T_5 = (((a, b), d), c)$	$\frac{1}{3}e^{-x} - \frac{5}{18}e^{-(x+y)}$	1	$\frac{1}{3}e^{-x} - \frac{1}{6}e^{-(x+y)} - \frac{1}{9}e^{-(3x+y)}$	1
$T_6 = (a, (b, (c, d)))$	$\frac{1}{3}e^{-y} - \frac{5}{18}e^{-(x+y)}$	1	$\frac{1}{18}e^{-(3x+y)}$	3
$T_7 = (b, (a, (c, d)))$	$\frac{1}{3}e^{-y} - \frac{5}{18}e^{-(x+y)}$	1	$\frac{1}{18}e^{-(3x+y)}$	3
$T_8 = (((a, c), b), d)$	$\frac{1}{18}e^{-(x+y)}$	2	$\frac{1}{3}e^{-y} - \frac{1}{3}e^{-(x+y)} + \frac{1}{18}e^{-(3x+y)}$	1
$T_9 = (((b, c), a), d)$	$\frac{1}{18}e^{-(x+y)}$	2	$\frac{1}{3}e^{-y} - \frac{1}{3}e^{-(x+y)} + \frac{1}{18}e^{-(3x+y)}$	1
$T_{10} = (((a, d), b), c)$	$\frac{1}{18}e^{-(x+y)}$	2	$\frac{1}{18}e^{-(3x+y)}$	3
$T_{11} = (((b, d), a), c)$	$\frac{1}{18}e^{-(x+y)}$	2	$\frac{1}{18}e^{-(3x+y)}$	3
$T_{12} = (((a, c), d), b)$	$\frac{1}{18}e^{-(x+y)}$	2	$\frac{1}{6}e^{-(x+y)} - \frac{1}{9}e^{-(3x+y)}$	2
$T_{13} = (((b, c), d), a)$	$\frac{1}{18}e^{-(x+y)}$	2	$\frac{1}{6}e^{-(x+y)} - \frac{1}{9}e^{-(3x+y)}$	2
$T_{14} = (((a, d), c), b)$	$\frac{1}{18}e^{-(x+y)}$	2	$\frac{1}{18}e^{-(3x+y)}$	3
$T_{15} = (((b, d), c), a)$	$\frac{1}{18}e^{-(x+y)}$	2	$\frac{1}{18}e^{-(3x+y)}$	3

y, (c, d): $x)$ or $(S_4, \lambda) = (((a, b)$: y, $c)$: x, $d)$, where x and y are the positive lengths in coalescent time units of the two internal branches (Fig. 2).

Our investigation of the consistency of the MDC criterion for symmetric and asymmetric species trees (S_1, λ) and (S_4, λ) makes use of the rearrangement inequality, which states that given two sequences of real numbers $a_1 \leq \cdots \leq a_n$ and $b_1 \geq \cdots \geq b_n$, the inequality

$$a_1 b_1 + \cdots + a_n b_n \leq a_{\pi(1)} b_1 + \cdots + a_{\pi(n)} b_n \tag{3.2}$$

holds for any permutation π of $\{1, \ldots, n\}$ (Hardy et al., 1934). Note that if a_1 is strictly smaller than each of $a_2, \ldots, a_n$ and b_1 is strictly greater than each of $b_2, \ldots, b_n$, then for any permutation π such that the equality in Eq. (3.2) holds, it is necessary that $\pi(1) = 1$ (and so $a_1 = a_{\pi(1)}$). For otherwise, let $\pi(1) = i > 1$, and let $\pi(j) = 1$ for some $j > 1$. Because $a_1 < a_i$ and $b_1 > b_j$, $(a_1 b_1 + a_i b_j) - (a_i b_1 + a_1 b_j) = (a_1 - a_i)(b_1 - b_j) < 0$. But this leads to a contradiction because the permutation π' in which $\pi'(1) = 1$, $\pi'(j) = i$, and $\pi'(k) = \pi(k)$ for $k \neq 1$ and $k \neq j$ produces a sum $\sum_{k=1}^{n} a_{\pi'(k)} b_k$ strictly smaller than the smallest sum $\sum_{k=1}^{n} a_k b_k$.

In the proof below, the rearrangement inequality is applied to the list of probabilities (considered as $\{b_i\}$) and the list of deep coalescence costs (considered as $\{a_i\}$) of the 15 gene trees. As observed in Section 2.2, if two species tree candidates S and S' have the same unlabeled topology, then $\{\alpha(T, S) \mid T \in R_1\} = \{\alpha(T, S') \mid T \in R_1\}$ and $\{\alpha(T, S) \mid T \in R_2\} = \{\alpha(T, S') \mid T \in R_2\}$. Therefore, the rearrangement inequality can be applied separately in R_1 and R_2 to the probabilities and deep coalescence costs of gene trees.

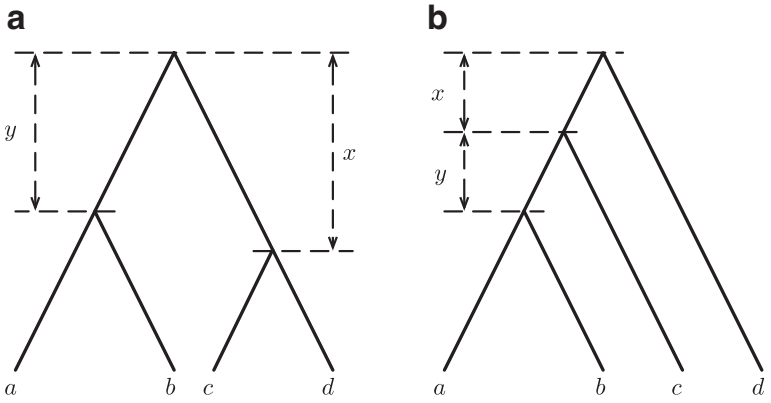

FIG. 2. Symmetric **(a)** and asymmetric **(b)** rooted, binary trees with leaf labels a, b, c, and d.

3.2.1. Symmetric species trees. Given that the true species tree is (S_1, λ), the probabilities of the 15 gene trees were computed by Rosenberg (2002), and they are reproduced in the second column of Table 1. The deep coalescence costs for reconciling each of the 15 gene trees within S_1 are also given in Table 1. It can be observed from the table that $\Pr(T_2 \mid S_1, \lambda) = \Pr(T_3 \mid S_1, \lambda)$, $\Pr(T_4 \mid S_1, \lambda) = \Pr(T_5 \mid S_1, \lambda)$, $\Pr(T_6 \mid S_1, \lambda) = \Pr(T_7 \mid S_1, \lambda)$, and $\Pr(T_8 \mid S_1, \lambda) = \cdots = \Pr(T_{15} \mid S_1, \lambda)$. Plugging the probability values and deep coalescence costs into Eq. (3.1), we have

$$\begin{aligned}\overline{\alpha}_{S_1,\lambda}(S_1) &= \sum_{i=1}^{15} \Pr(T_i \mid S_1, \lambda)\alpha(T_i, S_1) \\ &= \sum_{i=1}^{3} \Pr(T_i \mid S_1, \lambda)\alpha(T_i, S_1) + \sum_{i=4}^{15} \Pr(T_i \mid S_1, \lambda)\alpha(T_i, S_1) \qquad (3.3) \\ &= 4\Pr(T_2 \mid S_1, \lambda) + 2\Pr(T_4 \mid S_1, \lambda) + 2\Pr(T_6 \mid S_1, \lambda) + 16\Pr(T_8 \mid S_1, \lambda) \qquad (3.4)\end{aligned}$$

Let S' be a species tree candidate different from S_1. We aim to prove that $\overline{\alpha}_{S_1,\lambda}(S_1) < \overline{\alpha}_{S_1,\lambda}(S')$. There are two subcases to consider: S' is symmetric, and S' is asymmetric.

Tree S′ is symmetric. For gene trees in R_1, it can be seen seen from Table 1 that $\Pr(T_1 \mid S_1, \lambda) > \Pr(T_2 \mid S_1, \lambda) = \Pr(T_3 \mid S_1, \lambda)$. In fact, $\Pr(T_1 \mid S_1, \lambda)$ is the largest probability among the 15 probability values $\Pr(T_1 \mid S_1, \lambda), \ldots, \Pr(T_{15} \mid S_1, \lambda)$, so that the democratic vote method is consistent for symmetric species trees with four leaves (Degnan and Rosenberg, 2006). We also have $\alpha(T_1, S_1) = 0 < \alpha(T_2, S_1) = \alpha(T_3, S_1) = 2$. Moreover, because S_1 and S' have the same unlabeled topology, the list $\{\alpha(T_i, S'), i = 1, 2, 3\}$ is a permutation of $\{\alpha(T_i, S_1), i = 1, 2, 3\}$. Applying the rearrangement inequality to the lists $\{\alpha(T_i, S_1), i = 1, 2, 3\}$ and $\{\Pr(T_i \mid S_1, \lambda), i = 1, 2, 3\}$, we have

$$\sum_{i=1}^{3} \Pr(T_i \mid S_1, \lambda)\alpha(T_i, S_1) \le \sum_{i=1}^{3} \Pr(T_i \mid S_1, \lambda)\alpha(T_i, S'). \qquad (3.5)$$

For gene trees in R_2, one can check from Table 1 that $T_4, \ldots, T_7$ all have probabilities greater than those of $T_8, \ldots, T_{15}$, while their deep coalescence costs (for reconciling within S_1) are smaller. Trees S_1 and S' have the same unlabeled topology, and hence the list $\{\alpha(T_i, S'), i = 4, \ldots, 15\}$ is a permutation of $\{\alpha(T_i, S), i = 4, \ldots, 15\}$. We again apply the rearrangement inequality to the lists $\{\alpha(T_i, S_1), i = 4, \ldots, 15\}$ and $\{\Pr(T_i \mid S_1, \lambda), i = 4, \ldots, 15\}$, obtaining

$$\sum_{i=4}^{15} \Pr(T_i \mid S_1, \lambda)\alpha(T_i, S_1) \le \sum_{i=4}^{15} \Pr(T_i \mid S_1, \lambda)\alpha(T_i, S'). \qquad (3.6)$$

From Eqs. (3.3), (3.5), and (3.6), we have $\overline{\alpha}_{S_1,\lambda}(S_1) \le \overline{\alpha}_{S_1,\lambda}(S')$. Further, if $\overline{\alpha}_{S_1,\lambda}(S_1) = \overline{\alpha}_{S_1,\lambda}(S')$, then it is necessary that the equality in Eq. (3.5) holds. However, $\Pr(T_1 \mid S_1, \lambda)$ is strictly greater than $\Pr(T_2 \mid S_1, \lambda)$ and $\Pr(T_3 \mid S_1, \lambda)$, while $\alpha(T_1, S_1)$ is strictly smaller than $\alpha(T_2, S_1)$ and $\alpha(T_3, S_1)$. Using the equality condition in the rearrangement inequality, the equality in Eq. (3.5) holds only when $\alpha(T_1, S') = \alpha(T_1, S_1) = 0$, which is true only if $S' = S_1$. Therefore, $\overline{\alpha}_{S_1,\lambda}(S_1) < \overline{\alpha}_{S_1,\lambda}(S')$ for any symmetric species tree candidate $S' \neq S_1$.

Tree S′ is asymmetric. In this case, S' and S_4 have the same unlabeled topology, and so the lists $\{\alpha(T_i, S'), i = 1, 2, 3\}$ and $\{\alpha(T_i, S'), i = 4, \ldots, 15\}$ are permutations of $\{\alpha(T_i, S_4), i = 1, 2, 3\}$ and $\{\alpha(T_i, S_4), i = 4, \ldots, 15\}$, respectively. From Table 1, notice that $\alpha(T_1, S_4) = 1 < \alpha(T_2, S_4) = \alpha(T_3, S_4) = 2$. Applying the rearrangement inequality to the lists $\{\alpha(T_i, S_4), i = 1, 2, 3\}$ and $\{\Pr(T_i \mid S_1, \lambda), i = 1, 2, 3\}$, we have

$$\sum_{i=1}^{3} \Pr(T_i \mid S_1, \lambda)\alpha(T_i, S') \ge \Pr(T_1 \mid S_1, \lambda) + 4\Pr(T_2 \mid S_1, \lambda). \qquad (3.7)$$

For gene trees in R_2, the four smallest values among $\alpha(T_4, S_4), \ldots, \alpha(T_{15}, S_4)$ are smaller than or equal to 1, while the remaining eight values are at least 2. We have also noticed above that $T_4, \ldots, T_7$ all have probabilities greater than the probabilities of $T_8, \ldots, T_{15}$. However, the relative order between $\Pr(T_4 \mid S_1, \lambda) = \Pr(T_5 \mid S_1, \lambda)$ and $\Pr(T_6 \mid S_1, \lambda) = \Pr(T_7 \mid S_1, \lambda)$ depends on the branch lengths x and y. Assuming that $\Pr(T_4 \mid S_1, \lambda) \ge \Pr(T_6 \mid S_1, \lambda)$, then

$$\sum_{i=4}^{15} \Pr(T_i \mid S_1, \lambda)\alpha(T_i, S') \geq \Pr(T_4 \mid S_1, \lambda) + 2\Pr(T_6 \mid S_1, \lambda) + 22\Pr(T_8 \mid S_1, \lambda), \tag{3.8}$$

by applying the rearrangement inequality to the lists $\{\alpha(T_i, S_4), i = 4, \ldots, 15\}$ and $\{\Pr(T_i \mid S_1, \lambda), i = 4, \ldots, 15\}$.

From Eqs. (3.4), (3.7), and (3.8), we have

$$\begin{aligned}\overline{\alpha}_{S_1,\lambda}(S') - \overline{\alpha}_{S_1,\lambda}(S_1) &\geq \Pr(T_1 \mid S_1, \lambda) - \Pr(T_4 \mid S_1, \lambda) + 6\Pr(T_8 \mid S_1, \lambda) \\ &= 1 - e^{-x} - \frac{2}{3}e^{-y} + \frac{19}{18}e^{-(x+y)}.\end{aligned} \tag{3.9}$$

If we instead assume that $\Pr(T_6 \mid S_1, \lambda) > \Pr(T_4 \mid S_1, \lambda)$, then we obtain

$$\begin{aligned}\overline{\alpha}_{S_1,\lambda}(S') - \overline{\alpha}_{S_1,\lambda}(S_1) &\geq \Pr(T_1 \mid S_1, \lambda) - \Pr(T_6 \mid S_1, \lambda) + 6\Pr(T_8 \mid S_1, \lambda) \\ &= 1 - e^{-y} - \frac{2}{3}e^{-x} + \frac{19}{18}e^{-(x+y)}.\end{aligned} \tag{3.10}$$

It is straightforward to check that Eq. (3.9) and (3.10) are always greater than zero for all positive x and y. Consequently, $\overline{\alpha}_{S_1,\lambda}(S_1) < \overline{\alpha}_{S_1,\lambda}(S')$ for all asymmetric S'.

Whether the species tree candidate S' is asymmetric or symmetric, if $S' \neq S_1$, then $\overline{\alpha}_{S_1,\lambda}(S_1) < \overline{\alpha}_{S_1,\lambda}(S')$. Therefore, in the case of four-taxon, symmetric species trees, the unique asymptotic MDC tree matches the species tree topology. The MDC criterion is statistically consistent in this case.

3.2.2. Asymmetric species trees. Our treatment for the case of the asymmetric species tree $(S_4, \lambda) = (((a, b): y, c): x, d)$ is similar to the treatment for the symmetric species tree S_1 in Section 3.2.1. The probabilities of the 15 gene trees $T_1, \ldots, T_{15}$ given the species tree S_4, computed by Rosenberg (2002), are reproduced in the fourth column of Table 1. Again, there are two subcases to consider, depending on whether the species tree candidate S' is asymmetric or symmetric.

Tree S' is asymmetric. For gene trees in R_1, $\Pr(T_1 \mid S_4, \lambda) > \Pr(T_2 \mid S_4, \lambda) = \Pr(T_3 \mid S_4, \lambda)$, while $\alpha(T_1, S_4) = 1 < \alpha(T_2, S_4) = \alpha(T_3, S_4) = 2$. Also, because S' and S_4 have the same unlabeled tree topology, the list $\{\alpha(T_i, S'), i = 1, 2, 3\}$ is a permutation of $\{\alpha(T_i, S_4), i = 1, 2, 3\}$. By using the rearrangement inequality, we have

$$\sum_{i=1}^{3} \Pr(T_i \mid S_4, \lambda)\alpha(T_i, S_4) \leq \sum_{i=1}^{3} \Pr(T_i \mid S_4, \lambda)\alpha(T_i, S'). \tag{3.11}$$

For gene trees in R_2, we make three observations about their probabilities:

1. $\Pr(T_4 \mid S_4, \lambda) > \Pr(T_5 \mid S_4, \lambda) > \Pr(T_{12} \mid S_4, \lambda) = \Pr(T_{13} \mid S_4, \lambda) > e^{-(3x+y)}/18$;
2. $\Pr(T_4 \mid S_4, \lambda) > \Pr(T_8 \mid S_4, \lambda) = \Pr(T_9 \mid S_4, \lambda) > \Pr(T_{12} \mid S_4, \lambda)$; and
3. The remaining six trees in R_2—T_6, T_7, T_{10}, T_{11}, T_{14} and T_{15}—all have the smallest probability, $e^{-(3x+y)}/18$.

These observations can be easily verified from Table 1. For example, $\Pr(T_4 \mid S_4, \lambda) > \Pr(T_5 \mid S_4, \lambda)$ when

$$1 - e^{-x} - \frac{2}{3}e^{-y} + \frac{1}{2}e^{-(x+y)} + \frac{1}{6}e^{-(3x+y)} > 0,$$

which is equivalent to

$$\frac{2/3 - e^{-x}/2 - e^{-3x}/6}{1 - e^{-x}} < e^{y}.$$

However, $e^{y} > 1$ for positive y, while the left hand side is always smaller than 1 because

$$\frac{2}{3} - \frac{1}{2}e^{-x} - \frac{1}{6}e^{-3x} - (1 - e^{-x}) = -\frac{(e^{-x} - 1)^2(e^{-x} + 2)}{6},$$

which is smaller than zero for all positive x. We note that although the relative order of $\Pr(T_5 \mid S_4, \lambda)$ and $\Pr(T_8 \mid S_4, \lambda)$ depends on x and y, this is not important as the deep coalescence costs for reconciling either T_5 or T_8 within S_4 have the same value, 1.

As for the deep coalescence costs of gene trees in R_2, $\alpha(T_4, S_4) = 0 < \alpha(T_5, S_4) = \alpha(T_8, S_4) = \alpha(T_9, S_4) = 1 < \alpha(T_{12}, S_4) = \alpha(T_{13}, S_4) = 2$, while for each of the remaining six trees, the cost is 3. Based on the relative orders of the probabilities $\Pr(T_i \mid S_4, \lambda)$ and deep coalescence costs $\alpha(T_i, S_4)$, by using the rearrangement inequality, we have

$$\sum_{i=4}^{15} \Pr(T_i \mid S_4, \lambda)\alpha(T_i, S_4) \leq \sum_{i=4}^{15} \Pr(T_i \mid S_4, \lambda)\alpha(T_i, S'). \tag{3.12}$$

Equations (3.11) and (3.12) imply that $\overline{\alpha}_{S_4,\lambda}(S_4) \leq \overline{\alpha}_{S_4,\lambda}(S')$ for any asymmetric species tree candidate S'. If $\overline{\alpha}_{S_4,\lambda}(S_4) = \overline{\alpha}_{S_4,\lambda}(S')$, then equality must hold in Eq. (3.12). Because $\Pr(T_4 \mid S_4, \lambda)$ is strictly greater than the probabilities of other gene trees in R_2, while $\alpha(T_4, S_4)$ is strictly smaller than their deep coalescence costs, the equality in Eq. (3.12) holds only when $\alpha(T_4, S') = \alpha(T_4, S_4) = 0$, which in turn holds only when $S' = S_4$. Therefore, for any asymmetric species tree candidate $S' \neq S_4$, $\overline{\alpha}_{S_4,\lambda}(S_4) < \overline{\alpha}_{S_4,\lambda}(S')$.

Tree S' is symmetric. There are three symmetric species tree candidates—S_1, S_2, and S_3—and we consider each one of them in turn:

1. If $S' = S_1$, plugging the values of $\alpha(T_i, S_1)$ and $\Pr(T_i \mid S_4)$ into Eq. (3.1), we have

$$\overline{\alpha}_{S_4,\lambda}(S_1) = 1 - \frac{1}{3}e^{-x} + \frac{2}{3}e^{-y} + \frac{1}{6}e^{-(x+y)} + \frac{1}{18}e^{-(3x+y)}.$$

The expected deep coalescence cost of S_4, on the other hand, is

$$\overline{\alpha}_{S_4,\lambda}(S_4) = \frac{2}{3}(e^{-x} + e^{-y}) + \frac{1}{3}e^{-(x+y)} + \frac{5}{18}e^{-(3x+y)}.$$

Therefore, $\overline{\alpha}_{S_4,\lambda}(S_1)$ is smaller than $\overline{\alpha}_{S_4,\lambda}(S_4)$ when

$$1 - e^{-x} - \frac{1}{6}e^{-(x+y)} - \frac{2}{9}e^{-(3x+y)} < 0,$$

or equivalently, when

$$y < f(x) = \ln\left(\frac{3e^{2x} + 4}{18(e^{3x} - e^{2x})}\right). \tag{3.13}$$

2. If $S' = S_2$ or $S' = S_3$, then the expected deep coalescence cost of S' is

$$\overline{\alpha}_{S_4,\lambda}(S_2) = \overline{\alpha}_{S_4,\lambda}(S_3) = 2 - \frac{1}{3}e^{-y} - \frac{1}{6}e^{-(x+y)} + \frac{1}{18}e^{-(3x+y)},$$

which is smaller than $\overline{\alpha}_{S_4,\lambda}(S_4)$ if

$$2 - e^{-y} - \frac{2}{3}e^{-x} - \frac{1}{2}e^{-(x+y)} - \frac{2}{9}e^{-(3x+y)} < 0,$$

or, equivalently, if

$$y < g(x) = \ln\left(\frac{1}{2} + \frac{15e^{2x} + 4}{12(3e^{3x} - e^{2x})}\right). \tag{3.14}$$

Because we assume species tree branch lengths are positive, in order for a positive value of y to satisfy Eq. (3.13) or Eq. (3.14), the right hand sides of Eq. (3.13) and Eq. (3.14) must be positive. Both of these requirements yield the same condition:

$$18e^{3x} - 21e^{2x} - 4 < 0. \tag{3.15}$$

In addition, it is straightforward to verify that when Eq. (3.15) holds, $f(x) > g(x)$.

Figure 3 shows the plots of $f(x)$ and $g(x)$. To the right of the curve $f(x)$ in the figure, neither Eq. (3.13) nor Eq. (3.14) is satisfied, and the species tree S_4 is the asymptotic MDC tree. To the left of this curve, Eq. (3.13) holds, implying that $\alpha_{S_4,\lambda}(S_4)$ is not the smallest, and therefore, S_4 cannot be the asymptotic MDC tree. More precisely,

1. If $0 \le f(x) < y$ or if x is greater than or equal to the root $x_m \approx 0.2612$ of Eq. (3.15), then the asymptotic MDC tree is the species tree S_4;
2. If $0 \le g(x) < y < f(x)$, then S_1 is the only species tree candidate that has expected deep coalescence cost smaller than that of S_4;
3. If $0 < y \le g(x)$, then the trees S_1, S_2, and S_3 all have expected deep coalescence cost smaller than that of S_4.

We also note that in the boundary case where $y = f(x) > 0$, both S_1 and S_4 have the same expected deep coalescence cost. In the limit in which the number of gene trees goes to infinity, the MDC criterion considers S_1 and S_4 equally good, and so we also say that the MDC criterion is not statistically consistent in this case.

Because $f(x)$ approaches infinity as x approaches zero, for any given y, we can always make x small enough so that Eq. (3.13) holds. In other words, we can set x sufficiently small so that the species tree S_4 is not the asymptotic MDC tree. In addition, when $y < g(0) \approx 0.2559$, we can always choose sufficiently small x so that Eq. (3.14) holds, giving all three species tree candidates S_1, S_2, and S_3 smaller expected deep coalescence cost than the species tree S_4. Thus, for very small x, the MDC criterion is very likely to infer an incorrect estimate of the true species tree. The effects of x and y on the MDC criterion are similar to their effects on the democratic vote method (Degnan and Rosenberg, 2006).

Figure 3 also plots the anomaly zones of the democratic vote method, defined by functions $a(x) = \ln\left(\frac{2}{3} + \frac{3e^{2x}-2}{18(e^{3x}-e^{2x})}\right)$ and $b(x) = \ln\left(\frac{2}{3} + \frac{5e^{2x}-2}{6(3e^{3x}-2e^{2x})}\right)$ (Degnan and Rosenberg, 2006). Similar to the MDC criterion, the space of branch lengths x and y is also divided into three regions. To the right of the curve $a(x)$ in the figure, the democratic vote method is statistically consistent, that is, the most frequent gene tree has the same labeled topology as the species tree. In the region bounded by $a(x)$ and $b(x)$, there is exactly one labeled topology different from the species tree with higher probability than a matching gene tree. In the region below $b(x)$, there are three anomalous gene tree topologies.

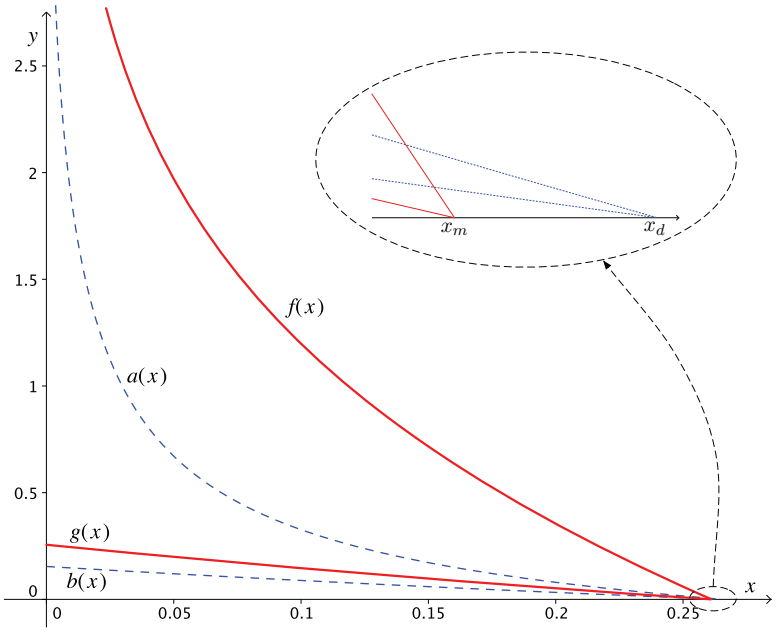

FIG. 3. Anomaly zones of the minimizing-deep-coalescence (MDC) criterion for asymmetric species trees with four leaves. In the region bounded by $y = g(x)$ and the two axes, there are three candidate species trees with lower expected deep coalescence costs than the true species tree, while in the region bounded by the y-axis, $y = g(x)$ and $y = f(x)$, there is one such anomalous candidate species tree. In this figure, the anomaly zones of the democratic vote method, defined by $a(x)$ and $b(x)$, are shown in dashed lines. The definitions of branch lengths x and y appear in Figure 2b.

It is interesting to see that the anomaly zones of the MDC criterion are larger than those of the democratic vote method. The largest value such that the MDC criterion is inconsistent when both branch lengths have the same value is $x = y \approx 0.2215$, whereas the corresponding length for the democratic vote is $x = y \approx 0.1569$ (Degnan and Rosenberg, 2006). However, it is not the case that $a(x)$ and $b(x)$ are always smaller than $f(x)$ and $g(x)$ (Fig. 3). The functions $a(x)$ and $b(x)$ intersect with the x-axis at $x_d \approx 0.2654$, while $f(x) = g(x) = 0$ at $x_m \approx 0.2612$. In the region bounded by the x-axis and the curves $f(x)$ and $a(x)$, the MDC criterion is consistent, whereas the democratic vote method is not consistent.

Remarks. It is possible to obtain the consistency properties of the MDC criterion on trees with four leaves by an exhaustive approach, that is, by directly computing the expected deep coalescence cost for every species tree candidate and comparing it with the corresponding cost of the parametric species tree. However, our approach using the rearrangement inequality provides a more concise proof that gives us some insight into the anomaly zones of the MDC criterion. As we have shown, only asymmetric species trees can produce anomalous candidate trees, which are symmetric. Intuitively, both for true species trees that are symmetric and for those that are asymmetric, the probabilities and deep coalescence costs of gene trees in each of the sets R_1 and R_2 are monotonic, but in opposite order. Therefore, by the rearrangement inequality, a candidate species tree that has the same unlabeled topology as the true species tree cannot have a smaller expected deep coalescence cost. This reasoning explains why the true species tree and anomalous candidate species trees must have different unlabeled topologies.

3.3. *Trees with five or more leaves*

We prove in this section that for any species tree topology with at least five leaves, there always exists a set of branch lengths that makes the MDC criterion infer the incorrect species tree topology in the limit as the number of sampled gene trees goes to infinity. Our approach is to make certain branches long enough to force the species tree to behave like an asymmetric four-leaf tree. For a rooted, binary (species) tree with at least five leaves, the longest path from its root to one of its leaves must have length at least three, for otherwise it cannot have more than four leaves. Call this path the "main" path in the tree, and consider the remaining parts of the tree as four subtrees $A_1, \ldots, A_4$ attached to this path (Fig. 4). None of these subtrees can be empty, although one or more of these subtrees (but not all) can each be a single leaf. If all internal branches of these subtrees, along with the four branches labeled $e_1, \ldots, e_4$, are arbitrarily long, then these subtrees can be "collapsed" into single leaves, resulting in an asymmetric four-leaf tree. Because we have already shown in Section 3.2.2 that such a species tree can mislead the MDC criterion, this reduction implies that the MDC criterion is also statistically inconsistent for trees with at least five leaves.

This argument can be made more rigorous as follows. If S has five or more leaves, then S has the same structure as the tree on the left in Figure 4, that is, S is obtained from $S_4 = (((a, b), c), d)$ by substituting leaves a, b, c, and d with nonempty subtrees $A_1, \ldots, A_4$, respectively. For each $i = 1, \ldots, 15$, let $S'_i = T'_i$ be the tree obtained from the corresponding $S_i = T_i$ in Table 1 using the same substitutions. Let h_i be a valid coalescent history for reconciling the four-leaf gene tree T_i within S_4. The coalescent history is a list of coalescence events along with a list of species tree internal branches (including the branch prior to the root of the species tree) on which they occur (Degnan and Salter, 2005; Rosenberg, 2007; Than et al., 2007).

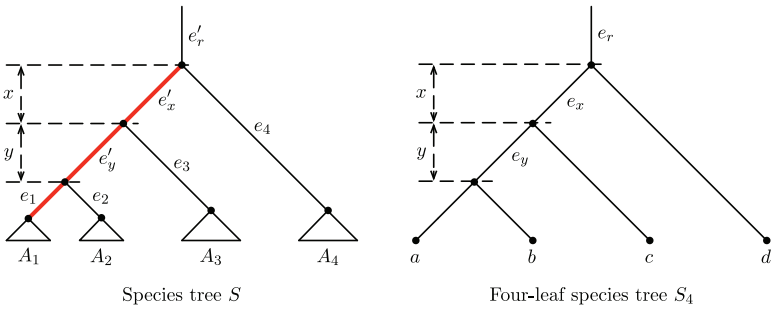

FIG. 4. A tree with at least five leaves **(left)**, illustrating the embedded structure of an asymmetric four-leaf tree **(right)**. The path with length at least three between the root and some leaf is shown with a thick line. Each of the triangles, representing taxon groups, contains at least one leaf.

From h_i, we create a coalescent history h_i' for reconciling the gene tree T_i' within S. We require that h_i' satisfy the following two conditions:

1. In each internal branch of subtrees $A_1, \ldots, A_4$, as well as in each of the branches labeled $e_1, \ldots, e_4$ in Figure 4, there is exactly one coalescence event. This implies that in each internal branch of $A_1, \ldots, A_4$ and in each branch $e_1, \ldots, e_4$, exactly two gene lineages enter, and they coalesce into one lineage.
2. Denoting the single gene lineages in $e_1, \ldots, e_4$ respectively as $g_1, \ldots, g_4$, we can think of them as gene lineages a, b, c, and d in S_4. We now require that the lineages $g_1, \ldots, g_4$ coalesce on the branches e_y', e_x', and e_r' of the species tree S in the same pattern as a, b, c, and d coalesce on the branches e_y, e_x, and e_r of S_4. Here, we attach the branches e_r' and e_r of infinite length to the roots of S and S_4 to accommodate coalescence events that occur prior to these root nodes.

It is easy to see that h_i' formed in this way is indeed a valid coalescent history for reconciling the gene tree T_i' within the species tree S. Let λ and λ' be the vectors of the lengths of branches of S_4 and S, repsectively. We give the branches e_x of S_4 and e_x' of S length x, and the branches e_y of S_4 and e_y' of S length y. We next claim that for any given $\epsilon > 0$, we can always make the internal branches of $A_1, \ldots, A_4$ and the branches $e_1, \ldots, e_4$ long enough so that

$$\rho \Pr(h_i) \le \Pr(h_i') \le \Pr(h_i), \tag{3.16}$$

where $\rho = (1-\epsilon)^4$, and $\Pr(h_i')$ and $\Pr(h_i)$ are the probabilities of coalescent histories h_i' and h_i, respectively.

The probability $\Pr(h_i')$ is the product of the probability for each internal branch of S and the branch e_r' that the coalescence events occur consistently with the gene tree T_i' (Degnan and Salter, 2005). In other words, if we denote by $p(e)$ the probability for the branch e that the coalescence events occurring along e are consistent with the gene tree T, then we can express $\Pr(h_i')$ as

$$\Pr(h_i') = \left(\prod_{k=1}^{4} p(e_k) \prod_{e \in \mathring{E}(A_k)} p(e) \right) p(e_y') p(e_x') p(e_r'),$$

where $\mathring{E}(A_k)$ is the set of the internal branches of the subtree A_k, $k = 1, \ldots, 4$; if A_k is a single-leaf subtree, then $\mathring{E}(A_k)$ is empty and we use the convention that $\prod_{e \in \mathring{E}(A_k)} p(e) = 1$. Because the probability for two lineages to coalesce on a branch e of length λ_e is $1 - \exp(-\lambda_e)$ (Hudson, 1983; Tajima, 1983), we have

$$p(e_k) \prod_{e \in \mathring{E}(A_k)} p(e) = (1 - \exp(-\lambda_{e_k})) \prod_{e \in \mathring{E}(A_k)} (1 - \exp(-\lambda_e)),$$

which approaches 1 as λ_{e_k} and λ_e, $e \in \mathring{E}(A_k)$, all approach infinity. Hence, for a given $\epsilon > 0$, we can always choose λ_{e_k} and λ_e, $e \in \mathring{E}(A_k)$, large enough so that

$$p(e_k) \prod_{e \in \mathring{E}(A_k)} p(e) \ge 1 - \epsilon.$$

We also have $p(e_k) \prod_{e \in \mathring{E}(A_k)} p(e) \le 1$ because all of its terms are probability values. Therefore,

$$(1-\epsilon)^4 p(e_y') p(e_x') p(e_r') \le \Pr(h_i') \le p(e_y') p(e_x') p(e_r').$$

Moreover, because our construction of h_i' requires that the gene lineages $g_1, \ldots, g_4$ coalesce on the branches e_y', e_x', and e_r' of S in the same pattern that lineages a, b, c, and d coalesce on the branches e_y, e_x, and e_r of S_4, we have

$$p(e_y') p(e_x') p(e_r') = \Pr(h_i),$$

and therefore,

$$\rho \Pr(h_i) = (1-\epsilon)^4 \Pr(h_i) \le \Pr(h_i') \le \Pr(h_i).$$

Our next step is to use Eq. (3.16) to derive the lower and upper bounds of the probability of observing the gene tree T_i' given the species tree (S, λ'). Let H_i be the set of *all* valid coalescent histories h_i for reconciling

the four-leaf gene tree T_i in the four-leaf species tree S_4, and let H'_i be the set of coalescent histories h'_i constructed from the corresponding h_i as described above. Because the probability of a gene tree is the sum of the probabilities of all valid coalescent histories for reconciling the gene tree within the species tree (Degnan and Salter, 2005), we have

$$\Pr(T'_i \mid S, \lambda') \geq \sum_{h'_i \in H'_i} \Pr(h'_i) \geq \sum_{h_i \in H_i} \rho \Pr(h_i) = \rho \Pr(T_i \mid S_4, \lambda), \tag{3.17}$$

On the other hand, the total probability of all coalescent histories for reconciling all gene trees in $R(X)$, discounting those in $H'_1, \ldots, H'_{15}$, satisfies

$$1 - \sum_{i=1}^{15} \sum_{h'_i \in H'_i} \Pr(h'_i) \leq 1 - \sum_{i=1}^{15} \sum_{h_i \in H_i} \rho \Pr(h_i) \leq 1 - \rho \sum_{i=1}^{15} \Pr(T_i \mid S_4, \lambda) = 1 - \rho.$$

Consequently, the probability of observing the gene tree T'_i given the species tree (S, λ') is bounded above by

$$\sum_{h'_i \in H'_i} \Pr(h'_i) + (1 - \rho) \leq \sum_{h_i \in H_i} \Pr(h_i) + (1 - \rho) = \Pr(T_i \mid S_4, \lambda) + (1 - \rho), \tag{3.18}$$

and the *total* probability of the $|R(X)| - 15$ gene trees in $R(X)$ other than $T'_1, \ldots, T'_{15}$ is bounded above by $1 - \rho$.

The deep coalescence cost $\alpha(T'_i, S)$ can be derived directly from $\alpha(T_i, S_4)$. Recall that S and T'_i are obtained from S_4 and T_i by replacing a, b, c, and d with subtrees $A_1, \ldots, A_4$. From Eq. (2.2), it is easy to see that the numbers of extra lineages in each internal branch of $A_1, \ldots, A_4$ as well as in $e_1, \ldots, e_4$ are all zero. Hence, the deep coalescence cost $\alpha(T'_i, S)$ is the sum of the numbers of extra lineages in the remaining two internal branches of S, e'_x and e'_y. If t is a maximal subtree of T_i with respect to, say, the branch e_x of S_4, then the subtree t' obtained from t by substituting each of the leaves a, b, c and d present in t with $A_1, \ldots, A_4$ is a maximal subtree of T'_i with respect to the branch e'_x of the species tree S. The converse is also true, that is, if t' is a maximal subtree of T'_i with respect to e'_x, then t is a maximal subtree of T_i with respect to e_x. By using Eq. (2.2), we have $\mathrm{xl}_S(T'_i, e'_x) = \mathrm{xl}_{S_4}(T_i, e_x)$ and $\mathrm{xl}_S(T'_i, e'_y) = \mathrm{xl}_{S_4}(T_i, e_y)$, and therefore,

$$\begin{aligned} \alpha(T'_i, S) &= \mathrm{xl}_S(T'_i, e'_x) + \mathrm{xl}_S(T'_i, e'_y) \\ &= \mathrm{xl}_{S_4}(T_i, e_x) + \mathrm{xl}_{S_4}(T_i, e_y) \\ &= \alpha(T_i, S_4). \end{aligned} \tag{3.19}$$

The expected deep coalescence cost for the candidate species tree that has the same labeled topology as S satisfies

$$\begin{aligned} \overline{\alpha}_{S, \lambda'}(S) &\geq \sum_{i=1}^{15} \Pr(T'_i \mid S) \alpha(T'_i, S) \\ &= \sum_{i=1}^{15} \Pr(T'_i \mid S) \alpha(T_i, S_4) \quad \text{(by Eq. (3.19))} \\ &\geq \sum_{i=1}^{15} \rho \Pr(T_i \mid S_4) \alpha(T_i, S_4) \quad \text{(by Eq. (3.17))} \\ &= \rho \overline{\alpha}_{S_4, \lambda}(S_4). \end{aligned}$$

A loose upper bound on the deep coalescence cost for reconciling two arbitrary trees in $R(X)$ is $|X|^2$. This bound follows because there can be at most $|X| - 1$ extra lineages in a tree branch and there are exactly $|X| - 2$ internal branches in a tree in $R(X)$. Consider the species tree candidate $S'_1 = ((A_1, A_2), (A_3, A_4))$. Using the same argument employed for proving $\alpha(T'_i, S) = \alpha(T_i, S_4)$, we also have $\alpha(T'_i, S'_1) = \alpha(T_i, S_1)$ for $i = 1, \ldots, 15$. Further, because $1 - \rho$ is the upper bound on the total probability of the $|R(X)| - 15$ gene trees in $R(X)$ other than $T'_1, \ldots, T'_{15}$, the expected deep coalescence cost for S'_1 satisfies

$$\begin{aligned}\overline{\alpha}_{S,\lambda'}(S_1') &\le \sum_{i=1}^{15} \Pr(T_i' \mid S, \lambda')\alpha(T_i', S_1') + (1-\rho)|X|^2 \\ &= \sum_{i=1}^{15} \Pr(T_i' \mid S, \lambda')\alpha(T_i, S_1) + (1-\rho)|X|^2 \\ &\le \sum_{i=1}^{15} (\Pr(T_i \mid S_4, \lambda) + (1-\rho))\alpha(T_i, S_1) + (1-\rho)|X|^2 \quad \text{(by Eq. (3.18))} \\ &\le \overline{\alpha}_{S_4,\lambda}(S_1) + (1-\rho)(|X|^2 + 30),\end{aligned}$$

where in the last step the term $30(1-\rho)$ arises because $\alpha(T_i, S_1) \le 2$ (Table 1). Therefore, if

$$\overline{\alpha}_{S_4,\lambda}(S_1) + (1-\rho)(|X|^2 + 30) < \rho\overline{\alpha}_{S_4,\lambda}(S_4),$$

which is equivalent to

$$\rho > \frac{\overline{\alpha}_{S_4,\lambda}(S_1) + (|X|^2 + 30)}{\overline{\alpha}_{S_4,\lambda}(S_4) + (|X|^2 + 30)}, \tag{3.20}$$

then $\overline{\alpha}_{S,\lambda'}(S_1') < \overline{\alpha}_{S,\lambda'}(S)$. As we have already shown in Section 3.2.2, when $y < f(x)$, as given in Eq. (3.13), $\overline{\alpha}_{S_4}(S_4)$ exceeds $\overline{\alpha}_{S_4}(S_1)$, making the right-hand side of Eq. (3.20) smaller than 1. Therefore, for ρ sufficiently close to 1 (i.e., for sufficiently small $\epsilon > 0$), the inequality in Eq. (3.20) is satisfied. This means that we can always assign appropriate lengths to the branches of S so that the species tree candidate with the smallest expected deep coalescence cost has a labeled topology different from the true species tree. Thus, in the limit as the number of genes tends to infinity, the MDC criterion will infer an incorrect species tree estimate.

Remarks. The techniques in this proof can also be used to simplify the proof of the inconsistency of the democratic vote method for trees with at least five leaves (Degnan and Rosenberg, 2006). Note that T_4' has the same labeled topology as the species tree S, and so in order to prove the inconsistency of the democratic vote method, we need to prove that $\Pr(T_4' \mid S, \lambda')$ is not the highest among the probabilities of all gene trees in $R(X)$. From Eq. (3.18), $\Pr(T_4' \mid S, \lambda') \le \Pr(T_4 \mid S_4, \lambda) + (1-\rho)$, while from Eq. (3.17), $\Pr(T_1' \mid S, \lambda') \ge \rho \Pr(T_1 \mid S_4, \lambda)$. Therefore, if

$$\rho \Pr(T_1 \mid S_4, \lambda) > \Pr(T_4 \mid S_4, \lambda) + (1-\rho)$$

or, equivalently, if

$$\rho > \frac{1 + \Pr(T_4 \mid S_4, \lambda)}{1 + \Pr(T_1 \mid S_4, \lambda)}, \tag{3.21}$$

then $\Pr(T_1' \mid S, \lambda') > \Pr(T_4' \mid S, \lambda')$. Because in the anomaly zone of the democratic vote method for asymmetric four-leaf trees (Eq. (4) in Degnan and Rosenberg, 2006), $\Pr(T_1 \mid S_4, \lambda) > \Pr(T_4 \mid S_4, \lambda)$, for ρ sufficiently close to 1, Eq. (3.21) holds. The inconsistency of the democratic vote method for species trees with at least five leaves immediately follows.

4. DISCUSSION

Although consistency properties of several methods for inferring species trees from gene trees have been investigated in a number of articles (Degnan and Rosenberg, 2006; Degnan et al., 2009; Liu et al., 2009, 2010; Mossel and Roch, 2010), no such results have been presented for species tree/gene tree reconciliation methods such as the MDC criterion. In this article, we have shown that the MDC criterion is inconsistent for asymmetric four-leaf species trees, and for species trees with at least five leaves. This result is interesting in that unlike other methods such as democratic vote or typical consensus methods, the MDC criterion is based on a perspective that specifically considers the mechanism of incomplete lineage sorting. However, it does not exploit all the elements of the multipsecies coalescent model, nor does it use all the information available in gene trees. In particular, the deep coalescence cost for reconciling a gene tree

within a species tree is used as an optimization criterion for finding an estimate of a species tree, and the probability of a gene tree given a species tree in the multispecies coalescent model is not used at all. Moreover, the lengths of the branches of gene trees are also not used. These facts might help to explain why the MDC criterion is not statistically consistent.

For species trees with three leaves or for symmetric species trees with four leaves, we have shown that the MDC criterion is statistically consistent. However, it is not statistically consistent for asymmetric four-leaf species trees, and we have obtained a complete characterization of the anomaly zones for the MDC criterion. There are three anomalous candidate species trees in the region below the curve $g(x)$ in Figure 3, and there is one such tree in the region bounded by $g(x)$ and $f(x)$. As for species trees with more than four leaves, we have provided an existence result that demonstrates the inconsistency of the MDC criterion. Future work will be required for characterizing the properties of the anomaly zones of the MDC criterion in full generality.

Simulation results demonstrate that it is more difficult for the MDC criterion to infer the correct estimate of the species tree topology in the case of recently diverged species than in the case of distantly diverged species (Maddison and Knowles, 2006; Than and Nakhleh, 2009). Our inconsistency result in this article provides a theoretical explanation for this phenomenon. In the case of asymmetric species trees with four leaves, we can see from Figure 3 that anomalous candidate species trees arise only when the branch x is quite short, less than approximately 0.2612 coalescent time units. In this case, the shapes of the anomaly zones for the MDC criterion are similar to those of the democratic vote method, although somewhat larger.

Finally, we have studied the consistency of the MDC criterion under the assumption that gene trees are known with certainty. This is an ideal case, and clearly, the accuracy of gene tree inference methods has an effect on the performance of the MDC criterion in practice. McCormack et al. (2009) have recently shown, through simulation studies, that the MDC criterion outperforms the maximum-likelihood method STEM (Kubatko et al., 2009) in certain cases for recently diverged species, while its performance is poorer than that of STEM for distantly diverged species. This result, along with the favorable performance of the MDC criterion on most examples considered by Than and Nakhleh (2009), suggests that despite its inconsistency, the MDC criterion might continue to be among the more desirable methods over large portions of the parameter space.

ACKNOWLEDGMENTS

Support was provided by the Burroughs Wellcome Fund, the Alfred P. Sloan Foundation, and the National Science Foundation (grant DEB-0716904).

DISCLOSURE STATEMENT

No competing financial interests exist.

REFERENCES

Bansal, M.S., Burleigh, J.G., and Eulenstein, O. 2010. Efficient genome-scale phylogenetic analysis under the duplication-loss and deep coalescence cost models. *BMC Bioinformatics* 11, S42.

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

Degnan, J.H., and Rosenberg, N.A. 2009. Gene tree discordance, phylogenetic inference and the multispecies coalescent. *Trends Ecol. Evol.* 24, 332–340.

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

Degnan, J.H., DeGiorgio, M,. Bryant, D., et al. 2009. Properties of consensus methods for inferring species trees from gene trees. *Syst. Biol.* 58, 35–54.

Hardy, G.H., Littlewood, J.E., and Polya, G. 1934. *Inequalities*. Cambridge University Press, New York.

Hudson, R.R. 1983. Testing the constant-rate neutral allele model with protein sequence data. *Evolution* 37, 203–217.

Kubatko, L.S., Carstens, B.C., and Knowles, L.L. 2009. STEM: species tree estimation using maximum likelihood for gene trees under coalescence. *Bioinformatics* 25, 971–973.

Liu, L., Yu, L., Pearl, D.K., et al. 2009. Estimating species phylogenies using coalescence times among sequences. *Syst. Biol.* 58, 468–477.

Liu, L., Yu, L., and Pearl, D.K. 2010. Maximum tree: a consistent estimator of the species tree. *J. Math. Biol.* 60, 95–106.

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

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

McCormack, J.E., Huang, H., and Knowles, L.L. 2009. Maximum likelihood estimates of species trees: how accuracy of phylogenetic inference depends upon the divergence history and sampling design. *Syst. Biol.* 58, 501–508.

Mossel, E., and Roch, S. 2010. Incomplete lineage sorting: consistent phylogeny estimation from multiple loci. *IEEE Trans. Comput. Biol. Bioinformatics* 7, 166–171.

Nichols, R. 2001. Gene trees and species trees are not the same. *Trends Ecol. Evol.* 16, 358–364.

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

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

Rosenberg, N.A. 2007. Counting coalescent histories. *J. Comput. Biol.* 14, 360–377.

Tajima, F. 1983. Evolutionary relationship of DNA sequences in finite populations. *Genetics* 105, 437–460.

Than, C.V., and Nakhleh, L.K. 2009. Species tree inference by minimizing deep coalescences. *PLoS Comput. Biol.* 5, e1000501.

Than, C.V., Ruths, D., Innan, H., et al. 2007. Confounding factors in HGT detection: statistical error, coalescent effects, and multiple solutions. *J. Comput. Biol.* 14, 517–535.

Address correspondence to:
Dr. Cuong V. Than
Center for Computational Medicine and Bioinformatics
University of Michigan
Ann Arbor, MI 48109

E-mail: tvcuong@umich.edu

This article has been cited by:

1. James H. Degnan, Noah A. Rosenberg, Tanja Stadler. 2011. The probability distribution of ranked gene trees on a species tree. *Mathematical Biosciences* . [CrossRef]
2. Taoyang Wu, Louxin Zhang. 2011. Structural properties of the reconciliation space and their applications in enumerating nearly-optimal reconciliations between a gene tree and a species tree. *BMC Bioinformatics* **12**:Suppl 9, S7. [CrossRef]
3. Jimmy Yang, Tandy Warnow. 2011. Fast and accurate methods for phylogenomic analyses. *BMC Bioinformatics* **12**:Suppl 9, S4. [CrossRef]