Are the Duplication Cost and Robinson-Foulds Distance Equivalent?

YU ZHENG¹ and LOUXIN ZHANG^{1,2}

ABSTRACT

In the tree reconciliation approach for species tree inference, a tree that has the minimum reconciliation score for given gene trees is taken as an estimate of the species tree. The scoring models used in existing tree reconciliation methods include the duplication, mutation, and deep coalescence costs. Since existing inference methods all are heuristic, their performances are often evaluated by using the Robinson-Foulds (RF) distance between the true species trees and the estimates output on simulated multi-locus datasets. To better understand these methods, we study the relationships between the duplication cost and the RF distance. We prove that the gap between the duplication cost and the RF distance is unbounded, but the symmetric duplication cost is logarithmically equivalent to the RF distance. The relationships between other reconciliation costs and the RF distance are also investigated.

Key words: gene duplication and loss, incomplete lineage sorting, metric equivalence, Robinson-Foulds distance, species tree inference, tree reconciliation.

1. INTRODUCTION

WITH MORE AND MORE GENOMES BEING fully sequenced, species tree inferences from multilocus datasets have become one of the basic bioinformatics tasks in comparative and evolutionary biology (Dunn et al., 2008). Reconstructing a species tree from a multi locus dataset requires reconciling the phylogenetic information contained in the genes. The challenge for such an analysis is that the genes sampled from the same set of species often contain conflicting signals and thus have discordant gene trees. Although some of these conflicting signals can be due to systematic errors occurring in phylogenetic analysis, they often reflect gene-specific mutational events. During evolution, recombination, gene duplication, gene loss, incomplete lineage sorting, or horizontal gene transfer, all might occur in a gene family (Fitch, 1970; Goodman et al., 1979; Maddison, 1997; Pamilo and Nei, 1988).

Existing methods for inferring species trees from gene trees can be divided into two broad categories: parametric methods (Ané et al., 2007; Kubatko et al., 2009; Liu and Pearl, 2007) and tree reconciliation methods (Warnow, 2013). Parametric methods adopt gene duplication or incomplete lineage-sorting models and infer species trees using either likelihood or Bayesian framework. These approaches have strong statistical foundation, but they can be computationally expensive. Therefore, tree reconciliation

¹Department of Mathematics, and ²Graduate School for Integrative Sciences and Engineering, National University of Singapore, Singapore.

methods are often used in phylogenetic studies on a genomic scale (Burleigh et al., 2011; Katz et al., 2012; Sanderson and McMahon, 2007).

Gene tree reconciliation methods take a collection of gene trees and output the tree that has the smallest reconciliation score for the gene trees as an estimate of the species tree. The scoring models used in these methods include the duplication cost (that is, the number of duplications), the mutation cost (that is, the number of gene duplication and loss events), the deep coalescence cost (that is, the number of incomplete lineage sorting events), and the Robinson-Foulds (RF) distance (Bansal et al., 2010; Bayzid et al., 2012; Chang et al., 2013; Chaudhary et al., 2013; Than and Nakhleh, 2009; Wehe et al., 2008; Yu et al., 2011). Since finding the parsimony tree from gene trees is NP-hard for the reconciliation scoring models (Bryant, 1997; Ma et al., 2000; Zhang, 2011), existing methods all compute a species tree by navigating through the part of the species tree space guided by the reconciliation score. The performances of these heuristic methods are usually evaluated using the RF distances between true species trees and their estimate output on simulation datasets (Chaudhary et al., 2013; Than and Nakhleh, 2009; Yang and Warnow, 2011). Hence, uncovering the relationships between the RF distance and the reconciliation costs is important to design good gene tree reconciliation methods. Unfortunately, many of these relationships are incompletely known. In particular, the relationship between the RF distance and the duplication cost is rather elusive, as both are bounded above by a linear function in the number of taxa. In this work, we mainly investigate this relationship. We consider all the reconciliation costs as different metrics in species tree spaces.

The rest of this article is divided into three sections. The basic concepts, notions, and the various reconciliation costs are introduced in Section 2. Several results are presented in Section 3. We first show that the gap between the duplication cost and the RF distance is unbounded. On the other hand, the symmetric duplication cost and the RF distance are proved to be logarithmically equivalent. A similar relationship between the gene loss cost and the RF distance is obtained. We also examine the distribution of the sizes of terraces for these metrics. We conclude with suggestions for future work in Section 4.

2. CONCEPTS AND NOTIONS

2.1. Species trees

A species tree T over a set of species X is a rooted tree in which one node is designated as the root, all edges are oriented away from the root, and the species in X are one-to-one mapped to leaves. We use $V(T)$ and $E(T)$ to denote the sets of nodes and (directed) edges in T , respectively. For $u, v \in V(T)$, v is the parent of u and, equivalently, u is a child of v if $(v, u) \in E(T)$. A node is called a leaf if it does not have any child. In this article, we focus on binary species trees in which every nonleaf node has two children. We also use the following notation:

- $p(u)$ denotes the parent of a nonroot node u in a tree;
- $\text{Ch}(u)$ denotes the set of two children of a nonleaf u in a tree;
- $V_{\text{lf}}(T)$ denotes the set of leaves in T ;
- $V_{\text{it}}(T)$ denotes the set of internal (i.e., nonleaf) nodes in T ;
- $T(u)$ denotes the subtree rooted at $u \in V_{\text{it}}(T)$, which consists of u and its descendants of u ;
- $P_T(-, u)$ denotes the unique path from the root to u in T ;
- $T|_U$ denotes the subtrees induced by a subset $U \subseteq V$, whose sets of nodes and branches are $V' = \{v \in V \mid \text{lca}(U) \preceq v \preceq u, u \in U\}$ and $E \cap (V' \times V')$, respectively.

For $u, v \in V(T)$, v is an ancestor of u and, equivalently, u is a descendant of v , denoted by $v \prec_T u$, if v is in $P_T(-, u)$. We write $v \preceq_T u$ if the relationship is uncertain, but we know that $v = u$ or $v \prec_T u$. Therefore, it is not hard to see that v is a leaf (terminal node) if and only if $v \not\prec_T w$ for any $w \in V(T)$. For $U \subseteq V(T)$, $\text{lca}(U)$ denotes the latest common ancestor (lca) of the nodes in U .

2.2. The RF distance

Let X be a set of species, and let $\mathcal{S}(X)$ denote the set of all species tree over X . Consider $S \in \mathcal{S}(X)$. A species $x \in X$ is called the label of the corresponding leaf $u \in V_{\text{lf}}(S)$, written $l(u)$. Conversely, we define $l^{-1}(x) = u$. For $u \in V(S)$, we define:

$$C(u) = \{l(u) \mid u \in V_{\text{lf}}(S(u))\},$$

called the (species) cluster of u . Define

$$\mathcal{C}(S) = \{C(u) \mid u \in V_{\text{it}}(S)\}.$$

Consider two species trees S' and S'' over X . They are identical if and only $\mathcal{C}(S') = \mathcal{C}(S'')$. Therefore, the dissimilarity of S' and S'' can be measured by the number of clusters that are found in one but not in the other, that is,

$$C_{\text{rf}}(S', S'') \stackrel{\text{def}}{=} |\mathcal{C}(S')/\mathcal{C}(S'')| + |\mathcal{C}(S'')/\mathcal{C}(S')|, \quad (1)$$

called the RF distance (Semple and Steel, 2003). Since both S' and S'' have $|X| - 1$ internal nodes and their roots' cluster is X , The following inequalities hold:

$$2 \leq C_{\text{rf}}(S', S'') \leq 2(|X| - 2), \quad \text{if } S' \neq S''. \quad (2)$$

2.3. Tree reconciliation costs

Consider $S', S'' \in \mathcal{S}(X)$. Since each leaf in these trees has a unique label in X , there is a one-to-one map between $V_{\text{lf}}(S')$ and $V_{\text{lf}}(S'')$. Such a correspondence can be extended into a map from $V(S')$ to $V(S'')$, $\lambda_{S', S''}$, as:

$$\lambda_{S', S''}(u) = \begin{cases} l^{-1}(l(u)) & \text{if } u \in V_{\text{lf}}(S'), \\ \text{lca}(\{\lambda(v) \mid v \in \text{Ch}(u)\}) & \text{if } u \in V_{\text{it}}(S'). \end{cases} \quad (3)$$

The *duplication cost* (DPC), $C_{\text{dup}}(S', S'')$, is defined as $|\text{Dup}(S', S'')|$, where

$$\text{Dup}(S', S'') = \{u \in V_{\text{it}}(S') \mid \lambda_{S', S''}(u) = \lambda_{S', S''}(v) \text{ for some } v \in \text{Ch}(u)\}. \quad (4)$$

The nodes in $\text{Dup}(S', S'')$ are called duplication nodes.

Proposition 1. For $S', S'' \in \mathcal{S}(X)$, $1 \leq C_{\text{dup}}(S', S'') \leq |X| - 2$ if $S' \neq S''$.

Proof. Since $S' \neq S''$, there exists $u \in V(S'')$ having the following properties (Fig. 1):

- (i) $C_{S''}(u) \neq C_{S'}(v)$ for any $v \in V(S')$;
- (ii) for each descendant t of u , that is, $t \succ_{S''} u$, $C_{S''}(t) = C_{S'}(v)$ for some $v \in V(S')$.

Let $\text{Ch}(u) = \{u_1, u_2\}$. By the above assumption, $C_{S''}(u_i) = C_{S'}(t_i)$ for some $t_i \in V(S')$ for $i = 1, 2$. Since $C_{S''}(u)$ is not a cluster in S' , t_1 and t_2 must not be siblings in S' , that is, $p(t_1) \neq p(t_2)$. Let $\text{lca}(t_1, t_2) = v$ and $\text{Ch}(v) = \{v_1, v_2\}$ such that $v_1 \preceq_{S'} t_1$ and $v_2 \preceq_{S'} t_2$. Since t_1 and t_2 are not siblings, either $t_1 \neq v_1$ or $t_2 \neq v_2$. If $t_2 \neq v_2$, $\lambda_{S', S''}(v_2)$ is in $P_{S''}(-, u)$. If $t_1 \neq v_1$, $\lambda_{S', S''}(v_1)$ is in $P_{S''}(-, u)$. Thus, $\lambda_{S', S''}(v_1)$, $\lambda_{S', S''}(v_2)$, or both are in $P_{S''}(-, u)$, the path from $r(T'')$ to u .

If $\lambda_{S', S''}(v_1) \preceq_{S''} \lambda_{S', S''}(v_2)$, $\lambda_{S', S''}(v) = \lambda_{S', S''}(v_1)$. If $\lambda_{S', S''}(v_1) \succeq_{S''} \lambda_{S', S''}(v_2)$, $\lambda_{S', S''}(v) = \lambda_{S', S''}(v_2)$. Therefore, v is a duplication node and hence $C_{\text{dup}}(S', S'') \geq 1$.

FIG. 1. Illustration of the proof of Proposition 1. The cluster of $u \in S''$ is not found in S' , but the clusters of the descendants of u are all in S' . If $t_1, t_2 \in S'$ such that $\{C(t_1), C(t_2)\} = \{C(u') \mid u' \in \text{Ch}(u)\}$, then $\text{lca}(t_1, t_2)$ is a duplication node under $\lambda_{S', S''}$.

Let $x \in V_{\text{it}}(S')$ have the largest distance from the root of S' . Then, x must have two leaf children, x_1 and x_2 . By definition, $\lambda_{S',S''}(x_1) \neq \lambda_{S',S''}(x) \neq \lambda_{S',S''}(x_2)$. Hence, x is not a duplication node when S' is mapped onto S'' . Since S' has $|X| - 1$ internal nodes, S' has at most $|X| - 1 - 1$ duplication nodes. Therefore, $C_{\text{dup}}(S', S'') \leq |X| - 2$. ■

For $e' = (u', v') \in E(S')$, we use $\lambda_{S',S''}(e')$ to denote the path from $\lambda_{S',S''}(u)$ to $\lambda_{S',S''}(v)$. We further use $|\lambda_{S',S''}(u, v)|$ to denote the number of the non-end nodes in $\lambda_{S',S''}(e')$. $|\lambda_{S',S''}(u, v)| = 0$ if $\lambda_{S',S''}(u) = \lambda_{S',S''}(v)$. The *gene loss cost*, $C_{\text{loss}}(S', S'')$, is defined as:

$$C_{\text{loss}}(S', S'') \stackrel{\text{def}}{=} \sum_{(u, v) \in E(S')} (|\lambda_{S',S''}(u, v)| + \chi_u), \quad (5)$$

where $\chi_u = 1$ if u is a duplication node such that $|\lambda_{S',S''}(u, \hat{u})| \neq 0$ for some $\hat{u} \in \text{Ch}(u)$ and $\chi_u = 0$ otherwise.

If S' is a gene tree and S'' is a species tree, $C_{\text{dup}}(S', S'')$ and $C_{\text{loss}}(S', S'')$ are the numbers of gene duplications and losses occurring in the duplication history of the genes represented by the leaves in S' defined by the lca map (Goodman et al., 1979).

For $e'' = (u'', v'') \in E(S'')$, define:

$$\text{DC}(e'') = \{(x, y) \in E(S') \mid \lambda_{S',S''}(x) \prec_{S''} u'' \text{ and } \lambda_{S',S''}(y) \succ_{S''} v''\}.$$

Clearly, $e \in E(S')$ is in $\text{DC}(e'')$ if and only if e'' is an edge in $\lambda_{S',S''}(e)$. The *deep coalescence cost*, $C_{\text{dc}}(S', S'')$, is defined as:

$$C_{\text{dc}}(S', S'') = \sum_{e'' \in E(S'')} (|\text{DC}(e'')| - 1). \quad (6)$$

The deep coalescence cost $C_{\text{dc}}(S', S'')$ is introduced to measure the dissimilarity of S' and S'' that are caused by incomplete lineage sorting for a gene tree S' and a species tree S'' (Maddison, 1997).

Finally, we note that all the reconciliation costs introduced above are not a symmetric metric in general. Hence, the *symmetric duplication cost* (SDPC) of S' and S'' , defined as $C_{\text{dup}}(S', S'') + C_{\text{dup}}(S'', S')$, is used to study species tree inference (Ma et al., 2000).

3. RESULTS

3.1. Unbounded gap between the DPC and RF distance

For two different trees over the same set of species, both the duplication cost and half of the RF distance are in the same range from 1 to $n - 2$, where n is the number of species involved in the trees. Therefore, the following question naturally arises:

Are the DPC and RF distance metrically equivalent?

and, equivalently, does there exist $c_1 \geq c_2 > 0$ such that

$$c_2 C_{\text{rf}}(S_1, S_2) \leq C_{\text{dup}}(S_1, S_2) \leq c_1 C_{\text{rf}}(S_1, S_2)$$

for any $S_1, S_2 \in \mathcal{S}(X)$?

Consider $S_1, S_2 \in \mathcal{S}(X)$ such that $C_{\text{rf}}(S_1, S_2) = 2(|X| - 2)$, that is, $\mathcal{C}(S_1) \cap \mathcal{C}(S_2) = \{X\}$. By definition, the parent of two sibling leaves in S_1 is not a duplication node under λ_{S_1, S_2} . Since such an internal node always exists in S_1 , we have:

$$C_{\text{dup}}(S_1, S_2) \leq |X| - 2 = \frac{1}{2} C_{\text{rf}}(S_1, S_2). \quad (7)$$

However, surprisingly, there are many pairs of trees S_1 and S_2 in $\mathcal{S}(X)$ satisfying that $C_{\text{rf}}(S_1, S_2) = 2(|X| - 2)$ but $C_{\text{dup}}(S_1, S_2) = 1$, where $|X| \geq 3$. For example, we consider a class of complete binary species trees, as illustrated in Figure 2. Let $n = 2^k$ for some $k \geq 1$ and $X = \{1, 2, \dots, n\}$. We set T to be the

FIG. 2. Complete binary species trees A_n and B_n for $n = 2^3$. $C_{\text{dup}}(A_n, B_n) = 1$ but $C_{\text{rf}}(A_n, B_n) = 2(n - 2)$.

rooted complete binary tree with n unlabeled leaves. A_n denotes the species tree over X obtained from T by labeling the i -th leaf (counting from left to right) with i (Fig. 2). B_n denotes the species tree obtained from T by labeling the i -th leaf with $\phi(i)$ that is defined as:

$$\phi(i) = \begin{cases} (i+1)/2 & \text{if } i \text{ is odd,} \\ 2^{k-1} + i/2 & \text{if } i \text{ is even} \end{cases}$$

for $i = 1, 2, \dots, 2^k$.

Proposition 2. (1). $C_{\text{rf}}(A_n, B_n) = 2(n - 2)$.

(2). $C_{\text{dup}}(A_n, B_n) = 1$.

(3). The DPC and the RF distance are not equivalent.

Proof. (1). For any nonroot $v \in V_{\text{it}}(A_n)$, either $C(u) \subseteq \{1, 2, \dots, n/2\}$, or $C(u) \subseteq \{n/2 + 1, n/2 + 2, \dots, n\}$. In B_n , the $(2i - 1)$ th and $(2i)$ th leaves are siblings, labeled with i and $n/2 + i$, respectively, where $i \leq n/2$. Therefore, for $u \in V_{\text{it}}(B_n)$,

$$C(u) \cap \{1, 2, \dots, n/2\} \neq \emptyset, \quad C(u) \cap \{n/2 + 1, n/2 + 2, \dots, n\} \neq \emptyset.$$

Combining these two facts together, we conclude that only the roots of A_n and B_n have the same cluster X . Hence, $C_{\text{rf}}(A_n, B_n) = 2(n - 2)$, as both trees have $n - 1$ internal nodes.

(2). Consider $u \in V_{\text{it}}(A_n)$. Notice that $C(u) = \{i, i + 1, \dots, i + 2k - 1\}$ for some i and k . For the left and right children, u_1 and u_2 , of u ,

$$\begin{aligned} C(u_1) &= \{i, i + 1, \dots, i + k - 1\}, \\ C(u_2) &= \{i + k, i + k + 1, \dots, i + 2k - 1\}. \end{aligned}$$

If u is in the left subtree of A_n , then $C(u) \subseteq \{1, 2, \dots, n/2\}$ and

$$\lambda_{A_n, B_n}(u) = \text{lca}(\{j, n/2 + j \mid j \in C(u)\}).$$

If u is in the right subtree of A_n ,

$$\lambda_{A_n, B_n}(u) = \text{lca}(\{j, j - n/2 \mid j \in C(u)\}).$$

Hence, the root r of A_n and its two children are mapped to the root of B_n , implying that r is a duplication node.

If u is in the left subtree of A_n such that $C(u) \neq \{1, 2, \dots, n/2\}$, then,

$$\begin{aligned} \lambda_{A_n, B_n}(u) &= \text{lca}(\{j, n/2 + j \mid i \leq j \leq i + 2k - 1\}), \\ \lambda_{A_n, B_n}(u_1) &= \text{lca}(\{j, n/2 + j \mid i \leq j \leq i + k - 1\}), \\ \lambda_{A_n, B_n}(u_2) &= \text{lca}(\{j, n/2 + j \mid i + k \leq j \leq i + 2k - 1\}). \end{aligned}$$

This implies that u is not a duplication node. Similarly, all the nodes in the right subtree are not duplication nodes. Therefore, $C_{\text{dup}}(A_n, B_n) = 1$.

(3). This statement follows directly from (1) and (2). ■

Although $C_{\text{dup}}(A_n, B_n) = 1$, we have that $C_{\text{dup}}(B_n, A_n) = n/2 - 2$. However, for any fixed large $k > 0$, there exist G and S such that $C_{\text{dup}}(G, S) + C_{\text{dup}}(S, G) < \frac{1}{k}|X|$ when X is large enough. For example, we take an integer m and define

$$X = \{a_{11}, a_{12}, \dots, a_{1m}\} \cup \{a_{21}, a_{22}, \dots, a_{2k}\} \cup \{a_{m1}, a_{m2}, \dots, a_{mm}\}.$$

Consider the two trees G and S over X defined in Figure 3. It is easy to see that $C_{\text{rf}}(G, S) = 2(m^2 - 2)$. The lca map $\lambda_{G,S}$ maps different nodes in G_i to different nodes in S for each i . Furthermore, it maps the roots of $G_1, G_2, \dots, G_m$ and their ancestors to the root of S . Thus, $C_{\text{dup}}(G, S) = m - 1$. Similarly, we can also show that $C_{\text{dup}}(S, G) = m - 1$. Hence, for any $k < m$, we have $C_{\text{dup}}(G, S) + C_{\text{dup}}(S, G) = 2m - 2 < \frac{1}{k}C_{\text{rf}}(G, S)$.

This example shows that even the symmetric duplication cost (SDPC) is not equivalent to the RF distance in the species tree space. However, they are weakly equivalent.

3.2. Logarithmic equivalence of the SDPC and RF distance

Consider $S', S'' \in \mathcal{S}(X)$. For the sake of simplicity, we use the following notations:

- μ denotes the lca map from S' to S'' , that is, $\mu = \lambda_{S', S''}$;
- $\psi = \lambda_{S'', S'}$;
- $\text{Dup}_\mu(S', S'')$ denotes the set of duplication nodes in S' under μ .

Lemma 1. For $S', S'' \in \mathcal{S}(n)$ such that $C_{\text{rf}}(S', S'') = 2(|X| - 2)$, $\psi(V_{it}(S'')) \subseteq \text{Dup}_\mu(S', S'')$ and $\mu(V_{it}(S')) \subseteq \text{Dup}_\psi(S'', S')$.

Proof. We prove the first fact by mathematical induction. By symmetry, the second fact holds. Consider $u \in V_{it}(S'')$. It has two children, u_1 and u_2 . There are three possible cases.

Case 1. u_1 and u_2 are leaves.

Since S' and S'' do not have any common node cluster except X , the leaves, w_1 and w_2 , that have the same labels as u_1 and u_2 are not siblings in S' . By definition, $\mu(w_i) = u_i$ and $\psi(u_i) = w_i$ for $i = 1, 2$. Let $v = \text{lca}(w_1, w_2)$. By assumption, $\psi(u) = \text{lca}(\psi(u_1), \psi(u_2)) = \text{lca}(w_1, w_2) = v$. Assume that the children of v are v_1 and v_2 such that $v_i \preceq_{S'} w_i$, $i = 1, 2$. Since w_1 and w_2 are not siblings, we have that $v_1 \neq w_1$, $v_2 \neq w_2$, or both. Without loss of generality, we may assume that $v_1 \neq w_1$. Since $\mu(w_1) = u_1$, which is a leaf, and $\mu(v_1) \neq u_1$, $\mu(v_1)$ is either equal to u or an ancestor of u , that is, $\mu(v_1)$ is in the path $P_{S'}(-, u)$. If $v_2 = w_2$, $\mu(v) = \text{lca}\{\mu(v_1), \mu(v_2)\} = \mu(v_1)$ (Fig. 4A), as $\mu(v_2) (= u_2)$ is a child of u and hence a descendant of $\mu(v_1)$. If $v_2 \neq w_2$, $\mu(v_2)$ is in $P_{S'}(-, u)$ and

FIG. 3. Trees G and S over m^2 species satisfying that $C_{\text{rf}}(G, S) = 2(m^2 - 2)$, but $C_{\text{dup}}(S, G) = C_{\text{dup}}(G, S) = m - 1$.

FIG. 4. Illustration of the proof of Lemma 1 in Case 1. Here, u has two leaf children, u_1 and u_2 , $\psi(u) = v$, and $\psi(u_i) = w_i$ for $i = 1, 2$. **(A)** Here, w_2 is a child of $\text{lca}(w_1, w_2)$, but w_1 is not. In this case, $\mu(v) = \mu(v_1)$. **(B)** Both w_1 and w_2 are not the children of v . The children, v_1 and v_2 , are mapped into the path from the root to u under μ .

hence $\mu(v_1)$ and $\mu(v_2)$ are comparable under $\prec_{S''}$, as they are in the path from the root to u (Fig. 4B). If $\mu(v_2) \prec_{S''} \mu(v_1)$, $\mu(v) = \mu(v_2)$. If $\mu(v_1) \prec_{S''} \mu(v_2)$, $\mu(v) = \mu(v_1)$. Therefore, $\psi(u) = v \in \text{Dup}_\mu(S', S'')$.

Case 2. Only one of u_1 and u_2 is a leaf.

This case is similar to Case 1. Let $u_1 \in V_{\text{lf}}(S'')$ and $u_2 \in V_{\text{it}}(S'')$. Assume w_1 is the leaf corresponding to u_1 in S' , that is, $\psi(u_1) = w_1 \in V_{\text{lf}}(S')$ and assume that $\psi(u_2) \in \text{Dup}_\mu(S', S'')$. If w_1 is a descendant of $\psi(u_2)$, then $\psi(u) = \text{lca}\{\psi(u_2), w_1\} = \psi(u_2) \in \text{Dup}_\mu(S', S'')$. If w_1 is not a descendant of $\psi(u_2)$, that is, $w_1 \not\prec_{S'} \psi(u_2)$, we let $v = \text{lca}(w_1, \psi(u_2))$. Then, $v = \psi(u)$ and $\mu(v)$ is in $P_{S''}(-, u)$. Since S' and S'' do not have any common cluster except X , $\mu(\psi(u_2))$ is also in $P_{S''}(-, u)$. If w_1 is a child of v , then $\mu(v) = \text{lca}(w_1, \mu(\psi(u_2))) = \mu(\psi(u_2))$. If w_1 is not a child of v , we assume that v_1 is the child of v such that $w_1 \succ_{S'} v_1$. Then $\mu(v_1)$ is in the path $P_{S''}(-, u)$ and is comparable with $\mu(\psi(u_2))$. We obtain that $\mu(v) = \mu(v_1)$ if $\mu(v_1) \prec_{S''} \mu(\psi(u_2))$ and $\mu(v) = \mu(\psi(u_2))$ otherwise. Therefore, $\psi(u)$ is v , a duplication node under μ .

Case 3. Both u_1 and u_2 are not leaves.

Assume $\psi(u_i) = t_i \in \text{Dup}_\mu(S', S'')$ for $i = 1, 2$. If $\psi(u) = \psi(u_i)$ for some i , it is obvious that $\psi(u) \in \text{Dup}_\mu(S', S'')$.

If $\psi(u) \neq t_1$ and $\psi(u) \neq t_2$, we have that $\psi(u) = \text{lca}(t_1, t_2)$. Let $\text{lca}(t_1, t_2) = v$ and $\text{Ch}(v) = \{v_1, v_2\}$ (Fig. 1). We prove that v is a duplication node under μ by considering the following two cases.

Case 3.1. $t_i \neq v_i$ for $i = 1$, or 2 . Assume $t_1 \neq v_1$. Since t_1 and u_1 have different node clusters, $\mu(t_1)$ is an ancestor of u_1 and hence $\mu(v_1)$ is in the path $P_{S''}(-, u)$. Since $\mu(v) = \text{lca}(\mu(v_1), \mu(v_2))$, $\mu(v)$ must be in $P_{S''}(-, u)$ and thus is equal to $\mu(v_1)$ if $\mu(v_1) \preceq_{S''} \mu(v_2)$. It is equal to $\mu(v_2)$ if $\mu(v_2) \prec_{S''} \mu(v_1)$, as $\mu(v_2)$ is also in $P_{S''}(-, u)$ in this case.

Case 3.2. $t_1 = v_1$ and $t_2 = v_2$. Since $C_{\text{rf}}(S', S'') = 2(|X| - 2)$ and the roots of S' and S'' have the same species cluster, $C_{S''}(u_1) \subseteq C_{S''}(v_1)$ but $C_{S''}(u_1) \neq C_{S''}(v_1)$. Hence, $\mu(v_1)$ is in the path $P_{S''}(-, u)$. Similarly, $\mu(v_2)$ is also in $P_{S''}(-, u)$. This implies that they are comparable under $\prec_{S''}$. Hence, $\mu(v) = \text{lca}(\mu(v_1), \mu(v_2))$, equal to $\mu(v_1)$ if $\mu(v_1) \prec_{S''} \mu(v_2)$ and equal to $\mu(v_2)$ otherwise. Hence, $\psi(u) = v \in \text{Dup}_\mu(S', S'')$. ■

For $v \in V(S'')$, we use $\mu^{-1}(v)$ to denote the set of nodes that are mapped to v under μ . If v is the root of S'' , $S'|_{\mu^{-1}(v)}$ is a subtree rooted at the root of S' . If v is a nonroot internal node in S'' , $S'|_{\mu^{-1}(v)}$ is a union of disjoint subtrees of S' .

Lemma 2. Let $v \in V_{\text{it}}(S'')$ be a nonroot node and m duplications nodes be mapped into $P_{S''}(-, p(v))$. Then, $\mu^{-1}(v)$ is a union of $m + 1$ subtrees of S' at most.

Proof. Let v be a nonroot internal node in S'' . Assume that $\mu^{-1}(v)$ is a union of $T_1, T_2, \dots, T_k$. Let T_i be rooted at x_i , $i = 1, 2, \dots, k$. Then, x_i 's are incomparable, that is, $x_i \not\prec_{S'} x_j$ and $x_j \not\prec_{S'} x_i$ for $i \neq j$. This implies that $S'|_{\{x_i\}}$ is a tree with k leaves in which $k - 1$ internal nodes are binary and other internal nodes have only one child.

Let y be a binary internal node that has two children in $S'|_{\{x_i\}}$. Clearly, $y = \text{lca}(x', x'')$ for some $x', x'' \in \{x_i\}$. Assume that y_1 and y_2 are the children of y in S' such that $y_1 \prec_{S'} x'$ and $y_2 \prec_{S'} x''$. Since $\mu(x') = \mu(x'') = v$, both $\mu(y_1)$ and $\mu(y_2)$ are in the path $P_{S''}(-, p(v))$. This implies that $\mu(y_1)$ and $\mu(y_2)$ are comparable under $\prec_{S''}$. Therefore, we have either $\mu(y_1) \prec_{S''} \mu(y_2)$ or $\mu(y_2) \prec_{S''} \mu(y_1)$. If the former is true, $\mu(y) = \mu(y_1)$. If the latter is true, $\mu(y) = \mu(y_2)$. Thus, y is a duplication node under μ .

Since all the $k - 1$ binary internal nodes in $S'|_{\{x_i\}}$ are duplication nodes under μ , $k - 1 \leq m$. Therefore, $k \leq m + 1$. $\blacksquare$

Lemma 3. Let $S', S'' \in \mathcal{S}(X)$. If there exist nonroot internal nodes $u \in V(S')$ and $v \in V(S'')$ such that $C_{S'}(u) = C_{S''}(v)$. Let S'_u denote the tree obtained from S' by replacing $S'(u)$ with a single leaf and S''_v the tree obtained from S'' by replacing $S''(v)$ with a single leaf with the same label. Then,

$$\begin{aligned} C_{\text{dup}}(S', S'') &= C_{\text{dup}}(S'(u), S''(v)) + C_{\text{dup}}(S'_u, S''_v), \\ C_{\text{rf}}(S', S'') &= C_{\text{rf}}(S'(u), S''(v)) + C_{\text{rf}}(S'_u, S''_v). \end{aligned}$$

Theorem 1. For $S', S'' \in \mathcal{S}(X)$, $C_{\text{dup}}(S', S'') \times C_{\text{dup}}(S'', S') \geq \frac{1}{4} C_{\text{rf}}(S', S'')$.

Proof. By Lemma 3, we may just assume that only the root clusters of these two trees are identical, that is, $C_{\text{rf}}(S', S'') = 2(|X| - 2)$. Let μ and ψ be defined as above. Assume that

$$\text{Img}_{\mu}(S', S'') \stackrel{\text{def}}{=} \{\mu(u) : u \in V_{\text{it}}(S')\} = \{v_1, v_2, \dots, v_k\} \subseteq V_{\text{it}}(S'').$$

Since each v_i is a duplication node with respect to the lca map ψ from S'' to S' ,

$$k \leq C_{\text{dup}}(S'', S'). \quad (8)$$

We further assume that $\mu^{-1}(v_i)$ contains q_i nodes and d_i out of them are duplication nodes, for each $1 \leq i \leq k$. Note that $S'|_{\mu^{-1}(v_i)}$ is a union of subtrees of S' and every internal node in such a subtree is a duplication node with respect to μ .

Assuming that $S'|_{\mu^{-1}(v_i)}$ is a union of t_i subtrees of S' , $T_1, T_2, \dots, T_{t_i}$. Let T_j have b_j internal nodes, each of which is a duplication node. Since T_j is binary, it has exactly $b_j + 1$ leaves, each of which is not a duplication node. Therefore, we have:

$$d_i = b_1 + b_2 + \dots + b_{t_i},$$

and

$$q_i = \sum_{j=1}^{t_i} [b_j + (b_j + 1)] = 2d_i + t_i.$$

Since there are at most $C_{\text{dup}}(S', S'') - d_i$ duplication nodes that are mapped in the path $P_{S''}(l, p(v_i))$, by Lemma 2, we have that $t_i \leq C_{\text{dup}}(S', S'') - d_i + 1$ if v_i is a nonroot node in S'' , and it is trivial for the root of S'' for which $t_1 = 1$. This implies that

$$q_i \leq 2d_i + (C_{\text{dup}}(S', S'') - d_i + 1) = d_i + C_{\text{dup}}(S', S'') + 1. \quad (9)$$

Hence, by Inequalities (8) and (9),

$$\begin{aligned} |X| - 1 &= \sum_{1 \leq i \leq k} q_i \\ &\leq \sum_{1 \leq i \leq k} d_i + k C_{\text{dup}}(S', S'') + k \\ &\leq C_{\text{dup}}(S', S'') + C_{\text{dup}}(S'', S') + C_{\text{dup}}(S'', S') C_{\text{dup}}(S', S''). \end{aligned}$$

By Proposition 1,

$$\begin{aligned} C_{\text{dup}}(S', S'') &\geq 1, \\ C_{\text{dup}}(S'', S') &\geq 1, \end{aligned}$$

and thus

$$C_{\text{dup}}(S', S'')C_{\text{dup}}(S'', S') \geq C_{\text{dup}}(S', S'') + C_{\text{dup}}(S'', S') - 1. \quad (10)$$

Using Inequality (10), we obtain the following fact:

$$\begin{aligned} &2C_{\text{dup}}(S', S'')C_{\text{dup}}(S'', S') \\ &\geq C_{\text{dup}}(S', S'')C_{\text{dup}}(S'', S') + C_{\text{dup}}(S', S'') + C_{\text{dup}}(S'', S') - 1 \\ &\geq |X| - 2 \\ &= \frac{1}{2}C_{\text{rf}}(S', S''). \end{aligned} \quad \blacksquare$$

Theorem 2. For arbitrary species trees S', S'' over the same set of species,

$$\frac{1}{2}\log_2 C_{\text{rf}}(S', S'') \leq \log_2(C_{\text{dup}}(S', S'') + C_{\text{dup}}(S'', S')) \leq \log_2 C_{\text{rf}}(S', S'').$$

Therefore, the SDPC and the RF distance are logarithmically equivalent

Proof. By Theorem 1,

$$C_{\text{dup}}(S', S'') + C_{\text{dup}}(S'', S') \geq 2\sqrt{C_{\text{dup}}(S', S'')C_{\text{dup}}(S'', S')} \geq \sqrt{C_{\text{rf}}(S', S'')}$$

and thus

$$\log_2(C_{\text{dup}}(S', S'') + C_{\text{dup}}(S'', S')) \geq \frac{1}{2}\log_2 C_{\text{rf}}(S', S'').$$

On the other hand, by Inequality (7) and Lemma 3,

$$C_{\text{dup}}(S'', S') + C_{\text{dup}}(S', S'') \leq \frac{1}{2}C_{\text{rf}}(S', S'') + \frac{1}{2}C_{\text{rf}}(S', S'') = C_{\text{rf}}(S', S'').$$

This implies the following inequality:

$$\log_2(C_{\text{dup}}(S', S'') + C_{\text{dup}}(S'', S')) \leq \log_2 C_{\text{rf}}(S', S''). \quad \blacksquare$$

3.3. The relationships between the RF distance and the gene loss cost

Consider $S', S'' \in \mathcal{S}(X)$. Let $u \in V_{\text{it}}(S'')$ such that $C(u) \in \mathcal{C}(S'')/\mathcal{C}(S')$. Assume $\lambda_{S', S''}(u) = v \in V_{\text{it}}(S')$ and $\text{Ch}(v) = \{v_1, v_2\}$. We have that $C(v_i) \cap C(u)$ is not empty for $i = 1, 2$. Since $C(u) \neq C(v)$, $C(u)$ is a proper subset of $C(v)$. This implies that $\lambda_{S', S''}(v)$ is in $P(-, u)$. We consider the following two cases.

Case 1. Only one of v_1 and v_2 is mapped into $P(-, u)$. We assume that v_1 is mapped to a descendant of u , that is, $u \prec_{S'} \lambda_{S', S''}(v_1)$. Then, $(p(u), u)$ is an edge in $\lambda_{S', S''}(v, v_1)$. Since $\lambda_{S', S''}(v_2)$ is in $P(-, u)$, v_2 has a descendant $\hat{v}$ such that $(p(u), u)$ is an edge in $\lambda_{S', S''}(p(\hat{v}), \hat{v})$. Thus, there is one extra lineage in $(p(u), u)$ at least.

Case 2. Both v_1 and v_2 are mapped into $P(-, u)$. Similar to Case 1, v_i has a descendant $\hat{v}_i$ such that $(p(u), u)$ is an edge in $\lambda_{S', S''}(p(\hat{v}_i), \hat{v}_i)$ for $i = 1, 2$. Thus, there is one extra lineage in $(p(u), u)$ at least. This proves:

$$C_{\text{dc}}(S', S'') \geq |\mathcal{C}(S'')/\mathcal{C}(S')| = \frac{1}{2}C_{\text{rf}}(S', S'').$$

Conversely, if $|X| \geq 3$ and $C_{\text{rf}}(S', S'') = 2(|X| - 2)$, by theorem 6 in Than and Rosenberg (2013),

$$C_{\text{dc}}(S', S'') \leq \frac{1}{2}(|X| - 2)(|X| - 1) \leq (|X| - 2)^2 = \frac{1}{4}C_{\text{rf}}(S', S'')^2.$$

It is easy to see that the result in Lemma 3 is also true for the deep coalescence cost. Combining all these facts together, we have:

$$\frac{1}{2}C_{\text{rf}}(S', S'') \leq C_{\text{dc}}(S', S'') \leq \frac{1}{4}C_{\text{rf}}(S', S'')^2, \quad S', S'' \in \mathcal{S}(X). \quad (11)$$

This implies the logarithmic equivalence of the RF distance and the deep coalescence cost.

For any $S', S'' \in \mathcal{S}(X)$, $C_{\text{loss}}(S', S'') = C_{\text{dc}}(S', S'') + 2C_{\text{dup}}(S', S'')$, proved in Zhang (2011). By Inequality (11),

$$\frac{1}{2}C_{\text{rf}}(S', S'') \leq C_{\text{loss}}(S', S'').$$

On the other hand, by Inequalities (7) and (11) and Lemma 3,

$$C_{\text{loss}}(S', S'') \leq \frac{1}{4}C_{\text{rf}}(S', S'')^2 + C_{\text{rf}}(S', S'') \leq C_{\text{rf}}(S', S'')^2.$$

Therefore, the RF distance and the gene loss cost are also logarithmically equivalent.

3.4. The size distribution of terraces in species tree space

Since species tree space grows exponentially with the number of taxa, understanding the landscape of the tree space is important in designing good heuristic algorithms that navigate through parts of this space guided by the parsimony score of a tree to infer species trees. Given a metric defined in the species tree space, the set of species trees with the same metric score from a species tree is called a terrace (Sanderson et al., 2011).

Here, we examine the size distribution of the DPC and RF terraces. Formally, a terrace defined by a species tree S for the DPC score c is $\{T \in \mathcal{S}(X) \mid C_{\text{dup}}(T, S) = c\}$. Similarly, we can define the concept for the RF distance. We adopt the Furnas rank (Furnas, 1984) to display the size distributions of these terraces. It is a complete order $\prec_F$ in the species tree space defined as follows. For $S', S'' \in \mathcal{S}(X)$, $S' \prec_F S''$ if one of the following three conditions holds:

1. The left subtree S'_l of S' has fewer leaves than the left subtree S''_l of S'' .
2. $|S'_l| = |S''_l|$, but $S'_l \prec_F S''_l$.
3. S'_l and S''_l are identical, but $S'_r \prec_F S''_r$, where S'_r and S''_r are the right subtrees of S and S' , respectively.

Figure 5 displays the size distributions of terraces with respect to the DPC (left panel) and RF distance (right panel) for the species tree space over 10 species. We observe the following facts:

FIG. 5. The size distribution of terraces defined by the DPC and the RF distance. All 98 topologies of the species tree space over 10 species are listed in increasing Furnas rank along the X-axis. Each color bar indicates the portion of species trees in a terrace that have the same score (given inside a circle in the corresponding band) with the corresponding species tree.

FIG. 6. The size distribution of the terraces defined by the symmetric duplication cost (left panel) and the gene loss cost (right panel). Ninety-eight topologies of the species trees over 10 species are listed in increasing Furnas rank along the X-axis. Each color bar indicates the portion of species trees in a terrace.

- The two terraces defined by the two largest values contain over 95% of the trees for the RF distance, independent of the Furnas rank of the reference tree.
- In contrast, the terraces defined by the middle values of the DPC are large, whereas the terraces defined by the extreme DPC cost are small. The size of the terrace of the largest DPC value increases slightly as the Furnas rank of the reference tree increases.

Figure 6 displays the size distributions of terraces with respect to the SDPC and gene loss costs. Interestingly, these size distributions are quite similar to each other. They indicate that SDPC and gene loss costs are more sensitive to the topology of the trees than the DPC and the RF distance.

4. CONCLUSION

All the tree reconciliation methods have high accuracy in inferring true species trees from real gene trees data. However, recently, Chaudhary et al. (2013) demonstrated by simulation that the RF-based parsimony approach often outperforms the reconciliation methods that use the gene duplication or mutation costs. A recent study of Yang and Warnow (2011) also suggests that the duplication-cost-based approach has poorer performance at large. One of our results is that the DPC and the RF distances are not equivalent. This provides a plausible explanation for the facts just mentioned. The weak equivalence of the RF distance and the gene loss cost also gives a clue to why the mutation cost is desirable for inferring species trees from a collection of gene trees (Chaudhary et al., 2013; Yang and Warnow, 2011).

Mathematical properties of the tree reconciliation costs have been studied in recent years. The deep coalescence cost has the Pareto property (Lin et al., 2012), meaning that a node cluster that appears in every input gene tree must also appear in the parsimony species tree. Than and Rosenberg (2013) report the formulas for the number of trees that achieve the maximum deep coalescence cost for a fixed gene or species tree. Górecki and Eulenstein (2014) presented a linear time algorithm for identifying a tree that maximizes the deep coalescence cost for a fixed gene or species tree. Zhang (2011) studied the relationship among the duplication cost, gene loss cost, and deep coalescence cost. The relationships between the nearest neighbor interchange distance and the reconciliation costs are also investigated (Bansal et al., 2009; Luo et al., 2011; Than and Rosenberg, 2013). However, many of mathematical properties of the reconciliation costs are incompletely known. For example, the distributions of these costs remain unclear for the uniform and Yule-Harding distribution (Harding, 1971) of species trees. In contrast, the distribution of the RF distance is much more clear. It is asymptotically Poisson with different means for several underlying species tree distributions (Steel and Penny, 1993). Our preliminary studies on terraces for the DPC and RF distances indicate that the distribution of the DPC seems to be different from that of the RF distance. Therefore, the following problems are open for future research:

1. Is the distribution of the DPC (respectively, SDPC) asymptotically normal?
2. Is the distribution of the gene loss cost asymptotically normal?

Solutions to these problems are definitely informative for studying species tree inference.

ACKNOWLEDGMENTS

L.X.Z would like to thank Jian Chen for helpful discussions. The work was supported by the Singapore MOE (Tier-1 FRC grant R146-000-177-112).

AUTHOR DISCLOSURE STATEMENT

No competing financial interests exist.

REFERENCES

- Ané, C., Larget, B., Baum, D.A., et al. 2007. Bayesian estimation of concordance among gene trees. *Mol. Biol. Evol.* 24, 412–426.
- Bansal, M.S., Burleigh, J.G., Eulenstein, O., et al. 2010. Robinson-Foulds supertrees. *Alg. Mol. Biol.* 5.
- Bansal, M.S., Eulenstein, O., and Wehe, A. 2009. The gene-duplication problem: Near-linear time algorithms for NNI-based local searches. *IEEE-ACM Trans. Comput. Biol. Bioinform.* 6, 221–231.
- Bayzid, M., Mirarab, S., and Warnow, T. 2012. Inferring optimal species trees under gene duplication and loss, 250–261. In *Pacific Symposium on Biocomputing*. World Scientific, Singapore.
- Bryant, D. 1997. Building trees, hunting for trees, and comparing trees [PhD Thesis]. Department of Mathematics, University of Canterbury, New Zealand.
- Burleigh, J.G., Bansal, M.S., Eulenstein, O., et al. 2011. Genome-scale phylogenetics: inferring the plant tree of life from 18,896 gene trees. *Syst. Biol.* 60, 117–125.
- Chang, W.-C., Górecki, P., and Eulenstein, O. 2013. Exact solutions for species tree inference from discordant gene trees. *J. Bioinform. Comput. Biol.* 11, 5.
- Chaudhary, R., Burleigh, J., and Fernandez-Baca, D. 2013. Inferring species trees from incongruent multi-copy gene trees using the robinson-foulds distance. *Alg. Mol. Biol.* 8, 28.
- Dunn, C.W., Hejnol, A., Matus, D.Q., et al. 2008. Broad phylogenomic sampling improves resolution of the animal tree of life. *Nature* 452, 745–749.
- Fitch, W.M. 1970. Distinguishing homologous from analogous proteins. *Syst. Biol.* 19, 99–113.
- Furnas, G.W. 1984. The generation of random, binary unordered trees. *J. Classification* 1, 187–233.
- Goodman, M., Czelusniak, J., Moore, G.W., et al. 1979. Fitting the gene lineage into its species lineage, a parsimony strategy illustrated by cladograms constructed from globin sequences. *Syst. Zool.* 28, 132–163.
- Górecki, P., and Eulenstein, O. 2014. Maximizing deep coalescence cost. *IEEE/ACM Trans. Comput. Biol. Bioinform.* 11, 231–242.
- Harding, E. 1971. The probabilities of rooted tree-shapes generated by random bifurcation. *Adv. Applied Prob.* 3, 44–77.
- Katz, L.A., Grant, J.R., Parfrey, L.W., et al. 2012. Turning the crown upside down: gene tree parsimony roots the eukaryotic tree of life. *Syst. Biol.* 61, 653–660.
- 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.
- Lin, H.T., Burleigh, J.G., and Eulenstein, O. 2012. Consensus properties for the deep coalescence problem and their application for scalable tree search. *BMC Bioinformatics* 13 (Suppl. 10), S12.
- Liu, L., and Pearl, D.K. 2007. Species trees from gene trees: reconstructing Bayesian posterior distributions of a species phylogeny using estimated gene tree distributions. *Syst. Biol.* 56, 504–514.
- Luo, C.-W., Chen, M.-C., Chen, Y.-C., et al. 2011. Linear-time algorithms for the multiple gene duplication problems. *IEEE-ACM Trans. Comput. Biol. Bioinform.* 8, 260–265.
- Ma, B., Li, M., and Zhang, L.X. 2000. From gene trees to species trees. *SIAM J. Comput.* 30, 729–752.
- Maddison, W.P. 1997. Gene trees in species trees. *Syst. Biol.* 46, 523–536.
- Pamilo, P., and Nei, M. 1988. Relationships between gene trees and species trees. *Mol. Biol. Evol.* 5, 568–583.
- Sanderson, M.J., and McMahon, M.M. 2007. Inferring angiosperm phylogeny from est data with widespread gene duplication. *BMC Evol. Biol.* 7 (Suppl. 1), S3.
- Sanderson, M.J., McMahon, M.M., and Steel, M. 2011. Terraces in phylogenetic tree space. *Science* 333, 448–450.
- Semple, C., and Steel, M.A. 2003. *Phylogenetics*. Oxford University Press, Oxford, U.K.
- Steel, M.A., and Penny, D. 1993. Distributions of tree comparison metrics: some new results. *Syst. Biol.* 42, 126–141.
- Than, C., and Nakhleh, L. 2009. Species tree inference by minimizing deep coalescences. *PLoS Comput. Biol.* 5, e1000501.

- Than, C.V., and Rosenberg, N.A. 2013. Mathematical properties of the deep coalescence cost. *IEEE-ACM Trans. Comput. Biol. Bioinformatics* 10, 61–72.
- Warnow, T. 2013. Large-scale multiple sequence alignment and phylogeny estimation, 85–146. In Chauve, C., El-Mabrouk, N., and Tannier, E., eds. *Models and Algorithms for Genome Evolution*. Springer, New York.
- Wehe, A., Bansal, M.S., Burleigh, J.G., et al. 2008. Duptree: a program for large-scale phylogenetic analyses using gene tree parsimony. *Bioinformatics* 24, 1540–1541.
- Yang, J., and Warnow, T. 2011. Fast and accurate methods for phylogenomic analyses. *BMC Biol.* 12 (Suppl. 9), S4.
- Yu, Y., Warnow, T., and Nakhleh, L. 2011. Algorithms for MDC-based multi-locus phylogeny inference: beyond rooted binary gene trees on single alleles. *J. Comput. Biol.* 18, 1543–1559.
- Zhang, L. 2011. From gene trees to species trees II: Species tree inference by minimizing deep coalescence events. *IEEE-ACM Trans. Comput. Biol. Bioinform.* 8, 1685–1691.

Address correspondence to:

Louxin Zhang

Department of Mathematics

10 Lower Kent Ridge Road

National University of Singapore

Singapore 119076

Republic of Singapore

E-mail: matzlx@nus.edu.sg