ALGORITHMS FOR MOLECULAR BIOLOGY

RESEARCH Open Access

Inferring species trees from incongruent multi-copy gene trees using the Robinson-Foulds distance

Ruchi Chaudhary[1,2*], John Gordon Burleigh[2] and David Fernández-Baca[1]

Abstract

Background: Constructing species trees from multi-copy gene trees remains a challenging problem in phylogenetics. One difficulty is that the underlying genes can be incongruent due to evolutionary processes such as gene duplication and loss, deep coalescence, or lateral gene transfer. Gene tree estimation errors may further exacerbate the difficulties of species tree estimation.

Results: We present a new approach for inferring species trees from incongruent multi-copy gene trees that is based on a generalization of the Robinson-Foulds (RF) distance measure to multi-labeled trees (mul-trees). We prove that it is NP-hard to compute the RF distance between two mul-trees; however, it is easy to calculate this distance between a mul-tree and a singly-labeled species tree. Motivated by this, we formulate the RF problem for mul-trees (MulRF) as follows: Given a collection of multi-copy gene trees, find a singly-labeled species tree that minimizes the total RF distance from the input mul-trees. We develop and implement a fast SPR-based heuristic algorithm for the NP-hard MulRF problem.
We compare the performance of the MulRF method (available at http://genome.cs.iastate.edu/CBL/MulRF/) with several gene tree parsimony approaches using gene tree simulations that incorporate gene tree error, gene duplications and losses, and/or lateral transfer. The MulRF method produces more accurate species trees than gene tree parsimony approaches. We also demonstrate that the MulRF method infers in minutes a credible plant species tree from a collection of nearly 2,000 gene trees.

Conclusions: Our new phylogenetic inference method, based on a generalized RF distance, makes it possible to quickly estimate species trees from large genomic data sets. Since the MulRF method, unlike gene tree parsimony, is based on a generic tree distance measure, it is appealing for analyses of genomic data sets, in which many processes such as deep coalescence, recombination, gene duplication and losses as well as phylogenetic error may contribute to gene tree discord. In experiments, the MulRF method estimated species trees accurately and quickly, demonstrating MulRF as an efficient alternative approach for phylogenetic inference from large-scale genomic data sets.

Background

With the proliferation of next generation sequencing technologies, there is great interest in using large genomic data sets for phylogenetic inference. One challenge for such phylogenomic analyses is that the genes sampled from the same set of species often produce conflicting trees [1]. Some of the incongruence among trees may be due to errors in the phylogenetic analyses. Alternately, the discordance may reflect biological processes such as recombination, gene duplication, gene loss, deep coalescence, or lateral gene transfer (LGT) [1-6]. Thus, in order to construct phylogenetic hypotheses from genomic data, it is necessary to address the incongruence among gene trees. Furthermore, any method for such phylogenetic analyses also must be computationally tractable for extremely large genomic data sets.

Constructing species phylogenies from a collection of gene trees requires summarizing and reconciling the phylogenetic information contained in the genes. The majority of such species tree reconstruction methods reconcile

*Correspondence: ruchic@ufl.edu
[1]Department of Computer Science, Iowa State University, Ames, IA 50011, USA
[2]Department of Biology, University of Florida, Gainesville, FL 32611, USA

the gene tree and species tree topologies using an optimality criterion based on a specific evolutionary process, such as gene duplication and loss or deep coalescence. In this paper we consider the problem of constructing species tree from gene trees using a tree distance measure that is not based on a specific biological process. We evaluate how our method performs in gene tree simulation experiments and with a large genomic data set from plants.

Existing methods for inferring species trees from collections of gene trees can be divided into two broad categories: non-parametric methods based on gene tree parsimony (GTP), and parametric methods that use likelihood (e.g., [7,8]) or Bayesian frameworks (e.g., [9-11]). GTP methods take a collection of discordant gene trees and try to find the species tree that implies the fewest evolutionary events. GeneTree [12], DupTree [13], and DupLoss [14] seek to minimize the number of duplications or duplications and losses. GeneTree [12], Mesquite [1], PhyloNet [15], and the method of [14] minimize deep coalescence events. The Subtree Prune and Regraft (SPR) supertree method [16] is based on minimizing the number of LGT events, and thus, it also can be considered a GTP method. Some of these methods have fast and effective heuristics, enabling the analysis of very large data sets. However, errors in the gene trees can mislead GTP analyses [17-19]. Furthermore, in some cases GTP methods may be statistically inconsistent, even when the gene tree topologies are correct [20]. Parametric methods exist based on either coalescence [7,10] or gene duplication and loss models [8]. Although such approaches have a strong statistical foundation, they can be extremely computationally expensive.

While the existing methods differ widely in their details, with the exception of [9], they are based on assumptions about the specific biological cause of discordance among gene trees. For example, GTP methods based on a duplication and loss cost implicitly assume that the differences between a gene tree and the species tree are caused by either gene duplications or losses. This does not necessarily mean that these methods will fail when their assumptions are incorrect, but it suggests that it is important to explore a range of different objectives for reconciling gene trees.

We present a new approach for constructing a species tree from discordant multi-copy gene trees based on a generic, non-biological distance measure. Our distance measure generalizes the Robinson-Foulds (RF) distance measure to multi-labeled trees (mul-trees) or trees where multiple leaves can have the same label. Our method takes as input a collection of multi-copy gene trees (mul-trees) and finds a species tree at minimum RF distance to the input gene trees. Our contributions are as follows:

- We study the problem of computing the RF distance between two mul-trees, and show that it is NP-complete (Theorem 1).
- We formulate an RF problem for mul-trees (MulRF) that takes a collection of multi-copy gene trees as input and constructs a binary species tree that is at minimum RF distance from each input gene tree (Section The MulRF Problem). A key component of this approach is a simple and efficient technique to compute the RF distance between an input multi-copy gene tree and a *singly-labeled* species tree. (Note the contrast with the previously-mentioned NP-completeness result.)
- MulRF is an NP-hard problem, so heuristics are required to estimate solutions for large data sets. We provide a fast $\Theta(nmk)$-time algorithm for the MulRF problem, where n is the total number of distinct leaf labels in the input collection of gene trees, m is the sum of n and the number of gene sequences in a input gene tree (assuming for convenience that all the input gene trees are built on approximately the same number of gene sequences), and k is the number of input gene trees (Section Solving the MulRF problem).
- We implemented the MulRF heuristic and examined its performance on simulated gene tree data sets that incorporate gene tree error, gene duplication and loss, and/or lateral gene transfer and a data set of nearly 2000 plant gene trees (Section Experimental evaluation).

We note that there has been much recent work on mul-trees ranging from constructing strict and majority rule consensus mul-trees to deriving diameter bounds for various metrics on mul-trees (see, [21-26]). Further, various problems related to RF distance have received attention. The RF distance has been extended to increase its robustness without sacrificing polynomial time computability [27,28]. These methods appear to work well when both input trees are singly-labeled, but there are no direct extensions of them for mul-trees. Alternatively, RF distance has been used in the supertree method for singly-labeled input trees [29,30] and the maximum-likelihood supertree approach of [31]. Here, we use RF distance for building species trees from mul-trees, which allows us to incorporate a wealth of genomic data from multi-copy genes into phylogenetic inference.

Our heuristic algorithm for MulRF problem shares several core concepts with unrooted RF supertree (URF) algorithm of [30], but there are theoretical and practical differences. In particular, our local search heuristic of MulRF is based on the SPR operation, unlike the p-Edge Contract and Refine operation (p-ECR) used for URF [30].

Typically, the SPR operation is more effective in exploring the space of trees compared to p-ECR operation; this also enables the MulRF heuristic to run as a standalone application on the given gene trees, independent from the rooted RF heuristic of [29]. In contrast, the p-ECR-based URF algorithm of [30] uses the output of the rooted method of [29] as a starting tree.

We performed gene tree simulation experiments to evaluate the accuracy of our method by comparing it against the model species tree used to simulate the data. We compared the species trees constructed by MulRF and GTP methods that consider only duplication [13], duplication and loss [14], and only LGT [16] with the true species trees. Our simulated data sets were too large to analyze with parametric methods, so we were unable to compare MulRF with these approaches. For example, when we ran Phyldog [8] on a single 50-taxon, 400 gene data set using 4 cores, it did not converge on a species tree in 110 hours. In contrast, MulRF gave an answer within a few seconds.

In all experiments, MulRF produced trees that are more similar to the true species trees than those obtained by the three GTP methods. This suggests that MulRF may be more robust than GTP methods to complex processes of gene evolution, including LGT, and in the presence of gene tree error. Furthermore, our algorithm runs quickly on moderate-size data sets, finishing in under two minutes on data sets containing 300 gene trees evolved over 100 taxon species trees. This suggests it is scalable for large-scale phylogenomic analyses. Finally, we examined the performance of the MulRF method on an unpublished plant gene tree data set with nearly 2000 gene trees from 22 species. The resulting species tree from the MulRF method was largely consistent with published plant phylogenies.

Preliminaries

Let X be a finite set of *labels*. A *phylogenetic mul-tree on* X (or *mul-tree*, for short) is a pair $\mathcal{T} = (T, \varphi)$ consisting of an unrooted tree T, whose leaf set is denoted by $\mathcal{L}(T)$, and where every internal vertex has degree at least three, along with a surjective map $\varphi : \mathcal{L}(T) \to X$. The tree T is called the *underlying tree* of $\mathcal{T}$ and φ is called the *labeling map* of $\mathcal{T}$. We say that $\mathcal{T}$ is a *singly-labeled tree* if φ is a bijection between $\mathcal{L}(T)$ and X (i.e., $|\varphi^{-1}(x)| = 1$ for all $x \in X$). Singly-labeled trees are also referred to as *phylogenetic X-trees* ([32]; page 17).

A mul-tree $\mathcal{T} = (T, \varphi)$ is *binary* if every internal vertex of T has degree 3. A vertex of T is said to be *unresolved* if its degree is greater than three. We use $V(T)$ and $E(T)$ to denote the set of vertices and the set of edges of T. The set of all internal vertices of T is $I(T) := V(T) \backslash \mathcal{L}(T)$. The *size* of $\mathcal{T}$, denoted by $|\mathcal{T}|$, is the number of elements in $\mathcal{L}(T)$.

Let $\mathcal{T} = (T, \varphi)$ be a mul-tree on X and $U \subseteq X$. Let $T[U]$ denotes the minimum subtree of T induced by the elements of $\{v \in \mathcal{L}(T) : \varphi(v) \in U\}$. The *restriction of* T *to* U, denoted $T_{|U}$ is the tree obtained from $T[U]$ by suppressing all vertices of degree two. The *restriction of* φ *to* U, denoted $\varphi_{|U}$ is the surjective mapping $\varphi_{|U} : \mathcal{L}(T_{|U}) \to U$, where for each $v \in \mathcal{L}(T_{|U}), \varphi_{|U}(v) = \varphi(v)$. The *restriction of* $\mathcal{T}$ *to* U, denoted by $\mathcal{T}_{|U}$, is the mul-tree on U given by $\mathcal{T}_{|U} = (T_{|U}, \varphi_{|U})$.

Two mul-trees $\mathcal{T}_1 = (T_1 = (V_1, E_1), \varphi_1)$ and $\mathcal{T}_2 = (T_2 = (V_2, E_2), \varphi_2)$ on X are isomorphic if there exists a bijection $\tau : V_1 \to V_2$, which induces a bijection between E_1 and E_2, subject to the condition that $\varphi_1(u) = \varphi_2(\tau(u))$ for all $u \in \mathcal{L}(T_1)$.

We define two basic operations on a mul-tree (T, φ). The *contraction* of an internal edge of T collapses that edge and identifies its two endpoints, yielding a new tree T' and a corresponding mul-tree (T', φ). (Note that, since T' and T have the same leaf sets, φ is also defined on T'.) Let v be an unresolved vertex of T. A *refinement* of v is obtained by partitioning the set of neighbors of v into two sets N_1 and N_2, such that $|N_1|, |N_2| > 1$, replacing v by two vertices v_1 and v_2 connected by an edge, and making the vertices of N_1 neighbors of v_1 and those in N_2 neighbors of v_2. This yields a new tree T', with the same leaf set as T, and a corresponding mul-tree (T', φ). Contraction and refinement can be viewed as inverses of each other (Figure 1).

Let $\mathcal{T}_1 = (T_1, \varphi_1)$ and $\mathcal{T}_2 = (T_2, \varphi_2)$ be mul-trees on X_1 and X_2, respectively, such that $X_1 \cap X_2 \neq \emptyset$. We say that $\mathcal{T}_1 = (T_1, \varphi_1)$ and $\mathcal{T}_2 = (T_2, \varphi_2)$ have *matching label multiplicities* if $|\varphi_1^{-1}(x)| = |\varphi_2^{-1}(x)|$ for all $x \in X_1 \cap X_2$. The *Robinson-Foulds (RF)* distance between two mul-trees $\mathcal{T}_1$ and $\mathcal{T}_2$ with identical label sets and matching label multiplicities, denoted by $RF(\mathcal{T}_1, \mathcal{T}_2)$, is defined as the minimum number of contractions and refinements

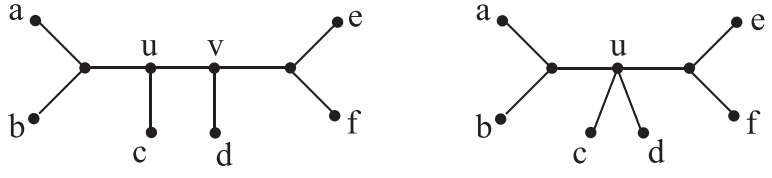

Figure 1 Contraction and refinement. Contracting edge {*u*,*v*} in the mul-tree on the left produces the mul-tree on the right. Conversely, refinement of vertex *u* in the mul-tree on the right produces the mul-tree on the left.

necessary to transform $\mathcal{T}_1$ into another mul-tree isomorphic to $\mathcal{T}_2$ [33,34]. (Note that [33] originally defined their distance measure for singly-labeled trees. Later on, [34] showed that the definition extends naturally to mul-trees.) We extend the RF distance to mul-trees $\mathcal{T}_1$, on X_1, and $\mathcal{T}_2$, on X_2, with $X_1 \subseteq X_2$ and matching label multiplicities, as $RF(\mathcal{T}_1, \mathcal{T}_2) := RF(\mathcal{T}_1, \mathcal{T}_{2|X_1})$.

Let $\mathcal{T} = (T, \varphi)$ be a mul-tree on X. Let M be a multiset on X such that the multiplicity in M of each element $x \in X$ is $|\varphi^{-1}(x)|$. A *split* $A|B$ of $\mathcal{T}$ is a bipartition of M, i.e., the sum of multiplicities of each element $x \in X$ in A and B is equal to the multiplicity of x in M. Multisets A and B are the *parts* of split $A|B$. (Note that if $\mathcal{T}$ is singly-labeled, then M, A, and B are sets.) The set of all splits induced by the internal edges of a mul-tree $\mathcal{T}$ is denoted by $\Sigma(\mathcal{T})$.

As Figure 2 illustrates, two mul-trees $\mathcal{T}_1$ and $\mathcal{T}_2$ such that $\Sigma(\mathcal{T}_1) = \Sigma(\mathcal{T}_2)$ may not be isomorphic. (See also ([34], Figure five) for a larger example.) On the other hand, by the Splits Equivalence Theorem ([32]; p. 44), if $\mathcal{T}_1$ and $\mathcal{T}_2$ are singly-labeled trees, then $\Sigma(\mathcal{T}_1) = \Sigma(\mathcal{T}_2)$ implies that $\mathcal{T}_1$ and $\mathcal{T}_2$ are isomorphic. Further, in this case, [33].

$$RF(\mathcal{T}_1, \mathcal{T}_2) = |(\Sigma(\mathcal{T}_1) \setminus \Sigma(\mathcal{T}_2)) \cup (\Sigma(\mathcal{T}_2) \setminus \Sigma(\mathcal{T}_1))| \quad (1)$$

Since mul-trees do not satisfy the Splits Equivalence Theorem, the RF distance between two of them cannot be computed by splits using expression (1). Nevertheless, as we show in Section The MulRF Problem, the formula will be useful for computing the RF distance between input gene tree and a species tree.

Ganapathy et al. [34] gave a worst-case exponential time algorithm for computing the RF distance between two mul-trees. The next result suggests that a polynomial time algorithm is unlikely.

Theorem 1. *Computing the RF distance between two mul-trees is NP-complete.*

Proof. See the Additional file 1. □

The MulRF Problem

A *profile* $\mathcal{P} = (\mathcal{T}_1, \mathcal{T}_2, \ldots, \mathcal{T}_k)$ is a tuple of mul-trees, also called *input mul-trees*, representing multi-copy gene trees, where, for each $i \in \{1, \ldots, k\}$, $\mathcal{T}_i$ has label set X_i. A *species tree* for $\mathcal{P}$ is a singly-labeled phylogenetic tree $\mathcal{S}$ on Y, where $Y = \bigcup_{i=1}^{k} X_i$.

A species tree $\mathcal{S}$ for $\mathcal{P}$ and a tree $\mathcal{T}$ in $\mathcal{P}$ will not, in general, have matching label multiplicities, since $\mathcal{S}$ is singly-labeled, while $\mathcal{T}$ need not be. In order to define $RF(\mathcal{T}, \mathcal{S})$, we will extend the species tree to add the missing duplicate leaf labels, thereby converting it into a mul-tree. We explain this formally next.

Let $\mathcal{T} = (T, \varphi)$ be an input mul-tree on X and $\mathcal{S} = (S, \phi)$ be a species tree on Y; thus, $X \subseteq Y$. The *extension of* $\mathcal{S}$ *relative to* $\mathcal{T}$ is a mul-tree $\mathcal{S}^* = (S^*, \phi^*)$ on Y, constructed from $\mathcal{S}$ by doing the following for each vertex $s \in \mathcal{L}(S)$ such that $|\varphi^{-1}(\phi(s))| > 1$. Let $k := |\varphi^{-1}(\phi(s))|$. Replace s by an internal vertex connecting to k leaves $\{l_1, \ldots, l_k\}$ labeled with $\phi(s)$; i.e., $\forall i (1 \leq i \leq k), \phi^*(l_i) = \phi(s)$. See Figure 3. We now define $RF(\mathcal{T}, \mathcal{S})$ to be $RF(\mathcal{T}, \mathcal{S}^*)$, where $\mathcal{S}^*$ is the extension of $\mathcal{S}$ relative to $\mathcal{T}$. We define the *RF distance from a profile* $\mathcal{P}$ *to a species tree* $\mathcal{S}$ *for* $\mathcal{P}$ as $RF(\mathcal{P}, \mathcal{S}) := \sum_{\mathcal{T} \in \mathcal{P}} RF(\mathcal{T}, \mathcal{S})$.

Let $\mathcal{B}(\mathcal{P})$ be the set of all binary species trees for $\mathcal{P}$.

Problem 1. (RF for MUL-Trees (MulRF))
Instance: A profile $\mathcal{P} = (\mathcal{T}_1, \mathcal{T}_2, \ldots, \mathcal{T}_k)$ *of mul-trees.*
Find: A species tree $\mathcal{S}^\star$ *for* $\mathcal{P}$ *such that* $RF(\mathcal{P}, \mathcal{S}^\star) = \min_{\mathcal{S} \in \mathcal{B}(\mathcal{P})} RF(\mathcal{P}, \mathcal{S})$.

Observe that the solution to the MulRF problem may not be unique. Further, the MulRF problem is NP-hard even when all the input mul-trees are singly labeled and their leaf label sets are identical [35]. Nevertheless, the "small" version of the problem —computing the RF distance between a profile of mul-trees and a species tree— is easy to solve. For each input mul-tree $\mathcal{T}$, we (i) construct the extension $\mathcal{S}^*$ of the species tree relative to $\mathcal{T}$; (ii) differentiate duplicate leaf labels in both $\mathcal{S}^*$ and $\mathcal{T}$; and (iii) apply the split-based formula (1) to compute the RF distance between the resulting singly-labeled phylogenetic trees. Next, we explain this process formally.

A *full differentiation* of a mul-tree $\mathcal{T} = (T, \varphi)$ on X is a singly-labeled tree $\mathbf{T} = (T, \varphi')$ on X' [34]. Note that both $\mathcal{T}$ and $\mathbf{T}$ have identical underlying trees, but the labeling map is surjective in the former, and bijective in the latter. Thus, X and X' may be different sets and $|X| \leq |X'|$. Intuitively, a full differentiation of a mul-tree differentiates the leaves that have identical leaf labels.

Let $\mathcal{T} = (T, \varphi)$ and $\mathcal{S} = (S, \phi)$ be two mul-trees, on X and Y, respectively, such that $\mathcal{T}$ and $\mathcal{S}$ have matching label multiplicities. Two full differentiations $\mathbf{T} = (T, \varphi')$ and

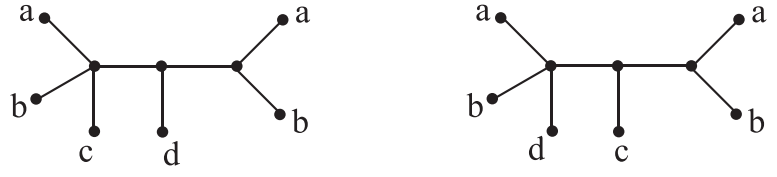

Figure 2 Contradicting example. Two mul-trees that induce the same set of splits but are not isomorphic. From ([23], Figure one).

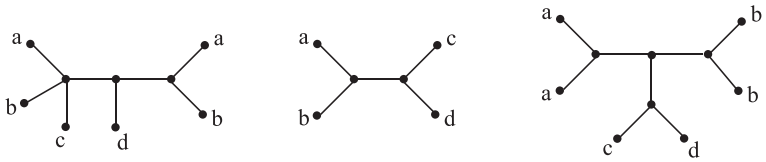

Figure 3 Rooting an unrooted tree. Phylogenetic tree **T** with leaf label set {*a, b, c, d, e*}. The rooted phylogenetic tree $\mathbb{T}$ with $r = a$ is also shown.

$\mathbf{S} = (S, \phi')$ of $\mathcal{T}$ and $\mathcal{S}$, respectively, are *consistent* if for each $a \in X \cap Y, \varphi'(\varphi^{-1}(a)) = \phi'(\phi^{-1}(a))$, i.e, both **T** and **S** have same set of new leaf labels for each common leaf label in $\mathcal{T}$ and $\mathcal{S}$. For instance, a consistent full differentiation can be obtained by relabeling each of the k copies of each leaf label a by $a_1, a_2, \ldots, a_k$ in both the mul-trees.

Theorem 2 ([34])**.** *Let $\mathcal{T} = (T, \varphi)$ and $\mathcal{S} = (S, \phi)$ be mul-trees with matching label multiplicities. Then, $RF(\mathcal{T}, \mathcal{S}) = \min\{RF(\mathbf{T}, \mathbf{S}) : \mathbf{T}$ and $\mathbf{S}$ are mutually consistent full differentiations of $\mathcal{T}$ and $\mathcal{S}$, respectively$\}$.*

We can prove the following result.

Theorem 3. *Let $\mathcal{T}$ be a mul-tree in a profile $\mathcal{P}$ and let $\mathcal{S}$ be a species tree for $\mathcal{P}$. Let $\mathcal{S}^*$ be the extension of $\mathcal{S}$ relative to $\mathcal{T}$. Then, for each pair of consistent full differentiations $(\mathbf{T}_1, \mathbf{S}_1)$ and $(\mathbf{T}_2, \mathbf{S}_2)$ of $\mathcal{T}$ and $\mathcal{S}^*$ we have $RF(\mathbf{T}_1, \mathbf{S}_1) = RF(\mathbf{T}_2, \mathbf{S}_2)$.*

Proof. Let $\mathcal{T} = (T, \varphi)$ be the input mul-tree on X. We prove the theorem by showing that for each $a \in X$, where $|\varphi^{-1}(a)| = k$, all $k!$ ways of uniquely relabeling corresponding k leaves in both $\mathcal{T}$ and $\mathcal{S}^*$ result into the same number of matched and unmatched splits in the corresponding mutually consistent full differentiations. The set of splits in $\mathcal{T}$ can be divided into two categories:

- *Category 1:* Splits that have all the leaves labeled with a in one part. Such a split will always have a match, irrespective of the labeling.
- *Category 2:* The remaining splits. Such splits are not present in $\mathcal{S}^*$, therefore, they will never have a match, irrespective of the labeling.

□

Thus, we can compute the RF distance between an input phylogenetic mul-tree and a species tree by computing the RF distance between *any* consistent full differentiations of the two trees. Since these full differentiations are singly-labeled trees, the RF distance between them can be found using Equation (1).

Solving the MulRF problem

Our local search heuristic for the MulRF problem starts with an initial (singly-labeled) species tree and explores the space of possible species trees in search of a *locally optimum* species tree, a species tree for $\mathcal{P}$ whose score is minimum within its "neighborhood". The neighborhood is defined in terms of the *Subtree Prune and Regraft* (*SPR*) operation [36]. An SPR operation on an unrooted, binary, singly-labeled phylogenetic tree $\mathcal{T} = (T, \varphi)$ on X cuts any edge $e \in E(T)$, thereby pruning a subtree t, and then regrafts t by the same cut edge to a new vertex obtained by subdividing a pre-existing edge in $T - t$ (Figure 4). The resulting phylogenetic tree is said to be *obtained from $\mathcal{T}$ by a single SPR operation*. The set of all phylogenetic trees obtained by the application of a single SPR operation on $\mathcal{T}$ is called the *SPR neighborhood* of $\mathcal{T}$, and is denoted by $SPR_{\mathcal{T}}$.

Problem 2. (**SPR Search**)
Instance: A profile $\mathcal{P} = (\mathcal{T}_1, \mathcal{T}_2, \ldots, \mathcal{T}_k)$ of mul-trees and a binary species tree $\mathcal{S}$ for $\mathcal{P}$.
Find: A species tree $\mathcal{S}^\star$ for $\mathcal{P}$ such that $\mathcal{S}^\star \in SPR_{\mathcal{S}}$ and $RF(\mathcal{P}, \mathcal{S}^\star) = \min_{\mathcal{S}' \in SPR_{\mathcal{S}}} RF(\mathcal{P}, \mathcal{S}')$.

The SPR Search based MulRF algorithm runs in two phases. In phase I, the algorithm quickly builds a likely suboptimal initial species tree using a greedy leaf adding procedure. This procedure first builds a phylogenetic tree on three randomly selected labels, and then it adds the remaining labels one at a time in a randomized order. In phase II, the algorithm performs a series of SPR Search iterations, each of which starts with an initial species tree and the input mul-trees. The output species tree of one SPR Search iteration serves as the initial species tree for the next iteration. When the resulting species tree of an SPR Search iteration is same as its initial species tree (i.e.,

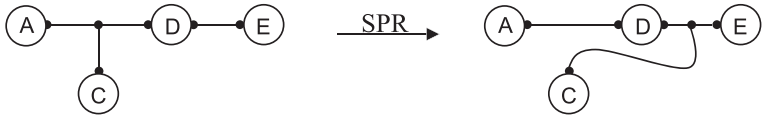

Figure 4 Species tree extension. From left to right, input mul-tree $\mathcal{T}$, the species tree $\mathcal{S}$, and the mul-tree $\mathcal{S}^*$. $\mathcal{S}^*$ is the extension of $\mathcal{S}$ relative to $\mathcal{T}$.

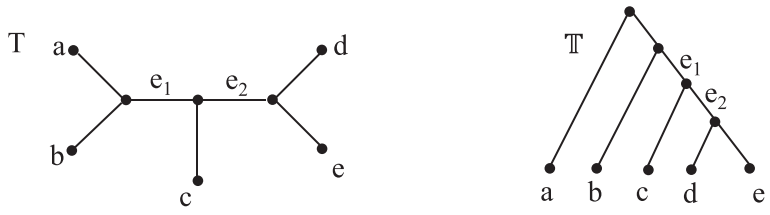

Figure 5 SPR operation. A schematic representation of the SPR operation.

there is no improvement in the score), the MulRF algorithm stops and returns the initial species tree of that iteration as the output.

Let the size of the input species tree $\mathcal{S}$ for the SPR Search problem be n, i.e. $n := |\mathcal{S}|$. For each $\mathcal{T} \in \mathcal{P}$, let $m := |\mathcal{T}| + |\mathcal{S}|$. (For convenience, we assume that all the input gene trees have approximately the same size.) Let k be the number of input gene trees in $\mathcal{P}$. In Section Solving the SPR search problem, we present an algorithm for the SPR search problem that runs in time $\Theta(nmk)$. (More precisely, if the size of the ith gene tree in the input profile be t_i then the complexity of our algorithm is $\Theta(\sum_{i=1}^{k} n(n+t_i))$. We made the assumption about the size of the input gene trees to simplify this complexity.) The algorithm relies on results from [30], which characterize the RF distance between unrooted phylogenetic trees in terms of least common ancestors in rooted versions of those phylogenies. These properties enable us to update the RF distance quickly after an SPR operation has been applied to one of the trees. For completeness, we briefly review these results in the next subsection. For a full discussion with proofs, see [30].

Robinson-Foulds distance and least common ancestors

In this subsection, we deal exclusively with singly-labeled trees, which we refer to simply as *phylogenetic trees*.

A phylogenetic tree $\mathbb{T} = (T, \varphi)$ is *rooted* if the underlying tree T is rooted; this means that T has exactly one distinguished vertex $rt(T)$, called the *root*. A rooted phylogenetic tree is *binary* if the root has degree two and every other internal vertex has degree three.

Let $\mathbb{T} = (T, \varphi)$ be a rooted phylogenetic tree on X. A vertex v of T is *internal* if $v \in V(T) \backslash (\mathcal{L}(T) \cup rt(T))$. The set of all internal vertices of T is denoted by $I(T)$. We define $\preceq_T$ to be the partial order on $V(T)$ where $x \preceq_T y$ if y is a vertex on the path from $rt(T)$ to x. If $\{x, y\} \in E(T)$ and $x \preceq_T y$, then y is the *parent* of x and x is a *child* of y. The *least common ancestor (LCA)* of a non-empty subset $L \subseteq V(T)$, denoted by $\mathrm{LCA}_T(L)$, is the unique smallest upper bound of L under $\preceq_T$.

For a rooted phylogenetic tree $\mathbb{T} = (T, \varphi)$ on X, let T_v denotes the subtree of T rooted at vertex $v \in V(T)$. For each node $v \in I(T)$, $C_{\mathbb{T}}(v)$ is defined to be the set of leaf labels $\{\varphi(u) \in X : u \in \mathcal{L}(T_v)\}$. Set $C_{\mathbb{T}}(v)$ is called a *cluster*. Let $\mathcal{H}(\mathbb{T})$ denote the set of all clusters of $\mathbb{T}$.

The RF distance between rooted phylogenetic trees $\mathbb{T} = (T, \varphi), \mathbb{S} = (S, \phi)$ on X, Y, respectively, such that $X = Y$, is defined as [33]

$$RF(\mathbb{T}, \mathbb{S}) := |(\mathcal{H}(\mathbb{T}) \backslash \mathcal{H}(\mathbb{S})) \cup (\mathcal{H}(\mathbb{S}) \backslash \mathcal{H}(\mathbb{T}))|.$$

When $X \subset Y$, we extend the RF distance in the same way as for unrooted trees. That is, $RF(\mathbb{T}, \mathbb{S}) := RF(\mathbb{T}, \mathbb{S}_{|X})$, where $\mathbb{S}_{|X} := (S_{|X}, \phi_{|X})$ is the rooted phylogenetic tree; here, $S_{|X}$ is obtained from $S[X]$ by suppressing all non-root degree-two vertices, $\phi_{|X}$ is the bijective mapping $\phi_{|X} : \mathcal{L}(S_{|X}) \to X$, where for each $v \in \mathcal{L}(S_{|X}), \phi_{|X}(v) = \phi(v)$.

Let $\mathbf{T}$ and $\mathbf{S}$ be two unrooted phylogenetic trees on X and Y, respectively, such that $X \subseteq Y$. Let $\mathbb{T}$ and $\mathbb{S}$ be the rooted phylogenetic trees that result from rooting the underlying trees of $\mathbf{T}$ and $\mathbf{S}$ at the branches incident on some arbitrarily-chosen but fixed leaf label $r \in X$ (Figure 5). (The leaf label sets of $\mathbb{T}$ and $\mathbb{S}$ are X and Y, respectively.)

Lemma 1 ([30])**.** *Let* $\mathbf{T}$ *and* $\mathbf{S}$ *be two unrooted phylogenetic trees on the same leaf label set, then* $RF(\mathbf{T}, \mathbf{S}) = RF(\mathbb{T}, \mathbb{S})$.

We now show how to compute the RF distance between $\mathbb{T} = (T, \varphi)$ on X and $\mathbb{S} = (S, \phi)$ on Y, when $X \subseteq Y$, without explicitly building $\mathbb{S}_{|X}$. We need two concepts. Let $v \in I(S)$. The *restriction* of $C_{\mathbb{S}}(v)$ to X is $\hat{C}_{\mathbb{T}}(v) := \{w \in Y : \phi^{-1}(w) \in \mathcal{L}(S_v) \text{and} w \in X\}$.

The *vertex function* $f_{\mathbb{S}}$ assigns each $u \in I(T)$ the value $f_{\mathbb{S}}(u) = |U|$, where $U := \{v \in I(S) : C_{\mathbb{T}}(u) = \hat{C}_{\mathbb{T}}(v)\}$. Observe that if $X = Y$, then for all $u \in I(T), f_{\mathbb{S}}(u) \leq 1$.

Lemma 2 ([30])**.** $RF(\mathbb{T}, \mathbb{S}) = |\mathcal{L}(T)| - |I(T)| + 2|\mathcal{F}_{\mathbb{S}}| - 2$, *where* $\mathcal{F}_{\mathbb{S}} := \{u \in I(T) : f_{\mathbb{S}}(u) = 0\}$.

We now describe a $O(n)$-time algorithm to compute the initial vertex function for $\mathbb{S}$ relative to $\mathbb{T}$, along with the RF distance between these two trees. The algorithm relies on LCAs. For $\mathbb{S}$ and $\mathbb{T}$, the *LCA mapping* $\mathcal{M}_{\mathbb{S},\mathbb{T}} : V(S) \to V(T) \cup \{\xi\}$ is defined as

$$\mathcal{M}_{\mathbb{S},\mathbb{T}}(u) := \begin{cases} \mathrm{LCA}_T(\phi^{-1}(\hat{C}_{\mathbb{T}}(u))), & \text{if } \hat{C}_{\mathbb{T}}(u) \neq \phi; \\ \xi, & \text{otherwise.} \end{cases}$$

See Figure 6.

Lemma 3 ([30]). *For all $u \in I(T), f_{\mathbb{S}}(u) = |B|$, where $B := \{v \in I(S) : \mathcal{M}_{\mathbb{S},\mathbb{T}}(v) = u$ and $|C_{\mathbb{T}}(u)| = |\hat{C}_{\mathbb{T}}(v)|\}$.*

The LCA computation of $\mathbb{T}$ is linear-time in the size of $\mathbb{T}$, and the LCA mapping from $\mathbb{S}$ to $\mathbb{T}$ can be done in $O(n)$ time [37] in bottom-up manner. Further, from Lemmas 2 and 3 we can compute the RF distance between $\mathbb{S}$ and $\mathbb{T}$ in $O(m)$ time as well.

Solving the SPR search problem

Let $\mathcal{T} = (T, \varphi)$ be a mul-tree (on X) in $\mathcal{P}$ and $\mathcal{S} = (S, \phi)$ be the input species tree (on Y). We now show how to compute the RF distance from $\mathcal{T}$ to each tree in the $\text{SPR}_{\mathcal{S}}$ neighborhood in time that is linear in the size of the neighborhood. By Theorem 3, computing the RF distance between $\mathcal{T}$ and each $\mathcal{S}' \in SPR_{\mathcal{S}}$ reduces to computing the RF distance between $\mathbf{T}$ and each $\mathbf{S}'$, where $\mathbf{T}$ and $\mathbf{S}'$ are the mutually consistent full differentiations of $\mathcal{T}$ and the extension of $\mathcal{S}$ relative to $\mathcal{T}$.

Suppose an SPR operation on $\mathcal{S}$ cuts the edge $e = \{x, y\} \in S$, and that $\dot{x}, \dot{y}$ are the subtrees of $S - e$ containing x, y, respectively. Suppose subtree $\dot{y}$ is pruned and regrafted by the same cut edge to a new vertex obtained by subdividing an edge in $\dot{x}$. The degree-two vertex x is suppressed and the new vertex is denoted by x. Observe that there are $O(n)$ possible edges in $\dot{x}$ to regraft $\dot{y}$. We perform regrafts in an order that leads to a constant time RF distance computation for each successive regraft.

We begin by regrafting $\dot{y}$ at an edge incident to a leaf in $\dot{x}$. Let $\overline{\mathcal{S}}$ be the phylogenetic tree obtained from performing the prune-and-regraft. Let $\mathbf{T}$ (on X') and $\overline{\mathbf{S}}$ (on Y') be the mutually consistent full differentiations of $\mathcal{T}$ and the extension of $\overline{\mathcal{S}}$. We compute the RF distance between $\mathbf{T}$ and $\overline{\mathbf{S}}$ using the algorithm described in Section Robinson–Foulds distance and least common ancestors. This method works by computing the RF distance between the rooted phylogenetic trees $\mathbb{T}$ and $\overline{\mathbb{S}}$ obtained by rooting $\mathbf{T}$ and $\overline{\mathbf{S}}$ at any leaf label in $X' \cap Y'$. (Note that if $X \cap \phi(\mathcal{L}(\dot{x})) = \emptyset$ or $X \cap \phi(\mathcal{L}(\dot{y})) = \emptyset$, then $\mathcal{T}$'s distance from $\overline{\mathcal{S}}$ is the same as its distance from $\mathcal{S}$.) The algorithm also computes the LCAs for $\mathbb{T}$ and the LCA mapping from $\overline{\mathbb{S}}$ to $\mathbb{T}$.

We perform the remaining regrafts of $\dot{y}$ on edges in $\dot{x}$ by iterating through the vertices of $\dot{x}$, starting from a leaf and exploring as far as possible along each branch before backtracking. The k^{th} regraft is performed on the edge between the k^{th} and $k + 1^{st}$ vertices in this iteration. Let us denote this ordering of edges by $\aleph$. See Figure 7. Observe that each two distinct consecutive edges in $\aleph$ are adjacent. We will show that, after the initial RF distance computation for $\overline{\mathcal{S}}$, we can compute in constant time the RF distance for the result of regrafting on each successive (adjacent) edges in $\aleph$.

Beginning with $\overline{\mathcal{S}}$, each $\mathcal{S}' \in SPR_{\mathcal{S}}$ helps in computing the RF distance of the next tree in the above regraft order. Assume that $\mathcal{S}' \in SPR_{\mathcal{S}}$ results from regrafting $\dot{y}$ at edge $\{a, b\}$ in $\dot{x}$, such that x subdivides the edge $\{a, b\}$ and neighbors to vertex y in $\dot{y}$, as shown in Figure 7. Let the rooted phylogenetic tree obtained after extending and differentiating $\mathcal{S}'$ be denoted by $\mathbb{S}'$. The LCA mapping and RF distance have been computed for $\mathbb{S}'$. Let $\mathcal{S}'' \in SPR_{\mathcal{S}}$ denote the tree obtained by regrafting $\dot{y}$ on edge $\{b, c\}$ in $\dot{x}$ and the rooted counterpart of $\mathcal{S}''$ is $\mathbb{S}''$.

Next, we find the vertices of $\mathbb{S}''$ whose LCA mappings have changed as a result of the SPR operation. Let T, S' and S'' be the underlying trees of $\mathbb{T}, \mathbb{S}'$ and $\mathbb{S}''$, respectively. Based on the topology of S', there are three cases:

1. *x is parent of b and b is parent of c.* For all $t \in I(S'')\backslash\{x, b\}, \mathcal{M}_{\mathbb{S}'',\mathbb{T}}(t) = \mathcal{M}_{\mathbb{S}',\mathbb{T}}(t)$. Further, $\mathcal{M}_{\mathbb{S}'',\mathbb{T}}(b) := \mathcal{M}_{\mathbb{S}',\mathbb{T}}(x)$, and $\mathcal{M}_{\mathbb{S}'',\mathbb{T}}(x) := \text{LCA}(\mathcal{M}_{\mathbb{S}',\mathbb{T}}(c), \mathcal{M}_{\mathbb{S}',\mathbb{T}}(y))$.
2. *b is parent of c and x.* For all $t \in I(S'')\backslash\{x\}, \mathcal{M}_{\mathbb{S}'',\mathbb{T}}(t) = \mathcal{M}_{\mathbb{S}',\mathbb{T}}(t)$. Further, $\mathcal{M}_{\mathbb{S}'',\mathbb{T}}(x) := \text{LCA}(\mathcal{M}_{\mathbb{S}',\mathbb{T}}(c), \mathcal{M}_{\mathbb{S}',\mathbb{T}}(y))$.
3. *b is parent of x and c is parent of b.* For all $t \in I(S'')\backslash\{b, x\}, \mathcal{M}_{\mathbb{S}'',\mathbb{T}}(t) = \mathcal{M}_{\mathbb{S}',\mathbb{T}}(t)$. Moreover, $\mathcal{M}_{\mathbb{S}'',\mathbb{T}}(x) := \mathcal{M}_{\mathbb{S}',\mathbb{T}}(b)$, and $\mathcal{M}_{\mathbb{S}'',\mathbb{T}}(b) := \text{LCA}(\mathcal{M}_{\mathbb{S}',\mathbb{T}}(d), \mathcal{M}_{\mathbb{S}',\mathbb{T}}(a))$.

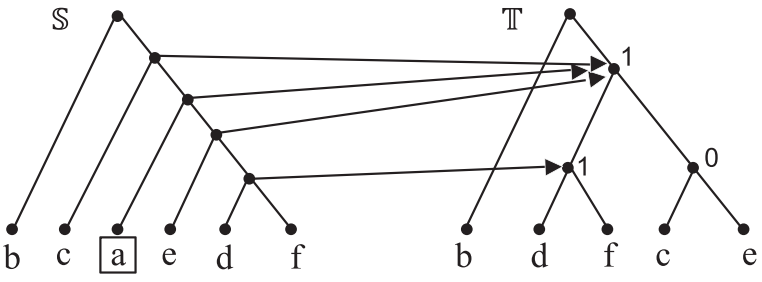

Figure 6 LCA mapping. The LCA mapping from $\mathbb{S}$ to $\mathbb{T}$. Vertex $\phi^{-1}(a)$ in the underlying tree of $\mathbb{S}$ is mapped to ξ as $a \notin X$. The internal vertices of the underlying tree of $\mathbb{T}$ are labeled with the values of the vertex function.

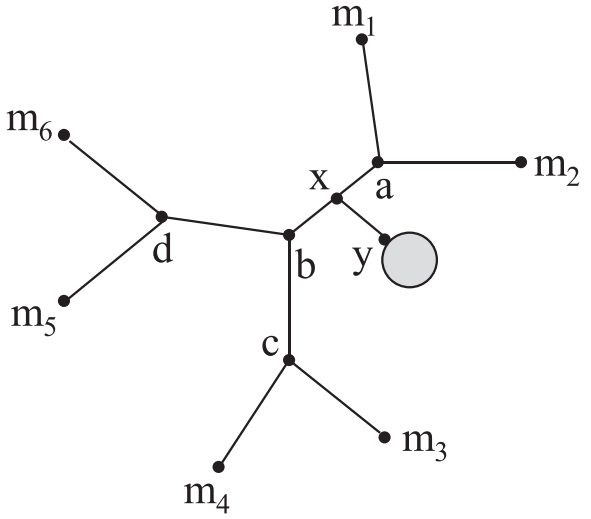

Figure 7 Iterating a tree. A phylogenetic tree with a subtree regrafted at an edge $\{a, b\}$ of the underlying tree. One iteration of vertices in this tree is $m_1, a, m_2, a, b, c, m_3, c, m_4, c, b, d, m_5, d, m_6, d, b, a, m_1$. The resulting ordering $\aleph$ is $\{m_1, a\}, \{a, m_2\}, \ldots, \{a, m_1\}$.

Since we can check in constant time which one of the above three cases holds, the LCA mappings can be updated in constant time too. Let H be a set $\{u \in I(T) : f_{\mathbb{S}''}(u) \neq f_{\mathbb{S}'}(u)\}$. Observe that H has at most four vertices, and thus it be computed in constant time. Let G denote the set $\{w \in H : f_{\mathbb{S}'}(w) = 0, \text{but } f_{\mathbb{S}''}(w) \geq 1\}$, and L denote the set $\{w \in H : f_{\mathbb{S}'}(w) \geq 1, \text{but } f_{\mathbb{S}''}(w) = 0\}$.

Lemma 4. $RF(\mathbb{S}'', \mathbb{T}) = RF(\mathbb{S}', \mathbb{T}) - 2|G| + 2|L|$.

Proof.

$$\begin{aligned} RF(\mathbb{S}'', \mathbb{T}) &= |\mathcal{L}(T)| - |I(T)| - 2 + 2|\mathcal{F}_{\mathbb{S}''}| \\ &= |\mathcal{L}(T)| - |I(T)| - 2 \\ &\quad + 2|\{u \in I(T) : f_{\mathbb{S}''}(u) = 0\}| \\ &= |\mathcal{L}(T)| - |I(T)| - 2 + 2|\mathcal{F}_{\mathbb{S}'}| \\ &\quad - 2|\{u \in H : f_{\mathbb{S}'}(u) = 0 \;\&\; f_{\mathbb{S}''}(u) \geq 1\}| \\ &\quad + 2|\{u \in H : f_{\mathbb{S}''}(u) = 0 \;\&\; f_{\mathbb{S}'}(u) \geq 1\}| \\ &= RF(\mathbb{S}', \mathbb{T}) - 2|G| + 2|L| \end{aligned}$$

□

Thus, after the initial regraft of $\dot{y}$ at a leaf in $\dot{x}$, we can compute in constant time the RF-distance between $\mathcal{T}$ and the species tree that results from each subsequent regraft.

Next, we present the results on complexity of our algorithm. Recall that n is the size of the species tree and m is the sum of n and the size of an input gene tree, where all the input gene trees are considered to have approximately the same size.

Lemma 5. *Let $\{x, y\}$ be an edge of S and let $\dot{x}$ and $\dot{y}$ be the subtrees of S containing x and y, respectively, that result from deleting $\{x, y\}$. The RF distance for the set of trees obtained by regrafting $\dot{x}$ (resp. $\dot{y}$) on each edge in $\dot{y}$ (resp. $\dot{x}$) can be computed in $\Theta(m)$ time.*

Proof. The RF distance computation for $\overline{\mathcal{S}}$, obtained by pruning $\dot{y}$ and regrafting at a leaf in $\dot{x}$, can be done in $\Theta(m)$ time. After $\overline{\mathcal{S}}$, the RF distance for each phylogenetic tree $\mathcal{S}'$ obtained by regrafting $\dot{y}$ on each edge in $\dot{x}$, can be computed in constant time by performing regrafts in the order of $\aleph$. There are $\Theta(n)$ edges in $\aleph$, thus the RF computation for all the phylogenetic trees can be done in $\Theta(m)$ time. The same argument applies for pruning $\dot{x}$ and regrafting on the edges in $\dot{y}$. □

Theorem 4. *The SPR Search problem can be solved in $\Theta(nmk)$ time.*

Proof. The underlying tree S of $\mathcal{S}$ has $\Theta(n)$ internal edges. For each edge $\{x, y\}$ in S, let $\dot{x}$ and $\dot{y}$ be the subtrees of S defined in the statement of Lemma 5. The RF distance for all the phylogenetic trees obtained by regrafting $\dot{x}$ (or $\dot{y}$) on each edge in $\dot{y}$ (or $\dot{x}$) can be computed in $\Theta(m)$ time from Lemma 5. Thus the RF distance for k input mul-trees can be checked in $\Theta(mk)$ time. The total execution time for $\Theta(n)$ internal edges must be $\Theta(nmk)$. □

Experimental evaluation

In order to evaluate the performance of the MulRF method, we implemented the heuristic algorithm of Section Solving the MulRF problem using C/C++. The MulRF software as well as simulated data sets (explained next) are freely available for download at http://genome.cs.iastate.edu/CBL/MulRF/.

Simulated data set

Methods. We used simulation experiments to evaluate the performance of MulRF and compare it to GTP methods. Since the MulRF method is designed for use with multi-copy gene trees, we focus on simulating genes that could have a history of duplication and/or lateral transfer. We first generated model species trees using the uniform speciation (Yule) module in the program Mesquite [38]. Two sets of model trees were generated: i) 50-taxon trees of height 220 thousand years (tyrs), ii) 100-taxon trees of height 440 tyrs. Note that the dates are relative; they do not have to represent thousands of years.

Next, we evolved 150 and 300 gene trees for each 50- and 100-taxon model species tree, respectively. For each gene tree a single gene birth node is chosen from the species tree nodes. Among all the simulated gene trees for a species tree, four gene trees have the gene birth node that is same as the root of the model tree. This represents

the sampling that would result from an experiment looking at genes from a few distantly related species. The rest had a gene birth node, which is selected at random using the model species tree topology and branch lengths. Starting from the children of the root, a Poisson process is tested along the parent edge of each node. If the birth occurs, the corresponding node becomes the birth node for that gene tree. This represents the sampling that would result from a study of closely related species.

We simulated the evolution of the gene trees within the model species tree using our C++ implementation of the duplication-loss model of [39]. We applied LGT events on the evolved gene trees, using the standard subtree transfer model of LGT. One LGT event causes the subtree rooted at a vertex c to be pruned and regrafted at an edge (a, b), where a and b together are not in the path from the root (of the tree) to c. We used gene duplication and loss (D/L) rate of 0.002 events/gene per tyrs and LGT rate of 0 to 2 events per gene tree. Note that the gene tree simulations without LGT follow a molecular clock model (equal rates of molecular evolution along all branches of the gene tree), but the simulations with LGT violate the molecular clock.

We generated gene trees based on four evolutionary scenarios: i) no duplications, losses, or LGT (called *none*), ii) D/L rate 0.002 and no LGT (called *dl*), iii) no duplication or loss, and LGT rate 2 (called *lgt*), and iv) D/L rate 0.002 and LGT rate 2 (called *both*). The parameter values (evolutionary scenario and model tree size) for each simulation are called the *model condition*; 20 model species trees were generated for each model condition. We deleted 0 to 25% of leaves (selected at random) from each gene tree to represent missing data or unsampled, which is common in almost all phylogenomic studies. For each gene tree, we used Seq-Gen [40] to simulate a DNA sequence alignment of length 500 based on the GTR+Gamma+I model. The parameters of the model were chosen with equal probability from the parameter sets estimated in [41] on three biological data sets, following [42]. We estimated maximum likelihood trees from each simulated sequence alignment using RAxML [43], performing searches from 5 different starting trees and saving the best tree. Since the true root of a gene tree with possible duplication and loss often is unknown, we rooted each estimated gene tree at the midpoint of the longest leaf-to-leaf path using Retree [44] before the species tree construction.

Species tree estimation. We estimated species trees with GTP minimizing only the number of duplications (Only-dup) [13], GTP minimizing duplications and losses (Dup-loss) [14], GTP minimizing LGT events (SPR supertree or SPRS for short) [16], and the MulRF heuristic. Both Only-dup and Dup-loss were executed with their default settings, including a fast leaf-adding heuristic for initial species tree construction. SPRS was run with 25 iterations of the global rearrangement search option. For 50-taxon data sets, it calculated the exact rSPR distance if it was 15 or less, and otherwise it estimated the rSPR distance using the 3-approximation. For the 100-taxon data sets, we used the 3-approximation of the rSPR distance. SPRS does not allow mul-trees as input. Therefore we only ran it on *none* and *lgt* data sets. Experiments were performed on the University of Florida High Performance Computing (HPC) cluster. We performed the experiments on the HPC cluster in order to simultaneously run the many simulations and phylogenetic analyses. However, all of the analyses (including SPRS, GTP, and MulRF) are sequential and easily run on a desktop machine. The running times are given in Table 1. The HPC cluster has cores of 2.3, 2.6, 2.9, or 2.66GHz on Opteron or Intel processors with 2 to 4GB RAM.

Results. We report the average topological error (ATE) for each model condition. This is the average of the normalized RF distance (dividing the RF distance by number of internal edges in both trees) between each of the 20 model species trees and their estimated species trees. An ATE of 0 indicates that two trees are identical, and an ATE of 100 indicates that two trees share no common splits.

For each set of 50- and 100-taxon model trees, the MulRF species trees are more accurate (lower ATE rate)

Table 1 Execution time

Num. Taxa	Sets	Only-dup	Dup-loss	SPRS	MulRF
50	*none*	< 1s	2 s	8 h 34 m 32 s	3 s
	lgt	< 1s	2 s	8 h 30 m 30 s	2 s
	dl	< 1s	3 s	NA	6 s
	both	< 1s	3 s	NA	6 s
100	*none*	9 s	37 s	21 h 34 m 25 s	58 s
	lgt	11 s	49 s	19 h 6 m 9 s	51s
	dl	9 s	30 s	NA	1 m 11 s
	both	11 s	37 s	NA	1 m 15 s

Running time for species tree estimations.

than those produced by the other three methods. The ATE rate of MulRF is 16.75% to 39.91% lower than the method of lowest ATE rate among other three methods (Figure 8).

In order to examine how Only-dup, Dup-loss, and SPRS methods perform when gene tree simulations only include events that these methods assume to be the source of discordance, we studied the performance of Only-dup and Dup-loss on evolutionary scenario *dl* and SPRS on *lgt*. In both evolutionary scenarios we found that the ATE rate of MulRF was lowest for both 50- and 100-taxon data sets (Figure 8). Surprisingly for *lgt*, while the ATE rate of SPRS was lower than Only-dup and Dup-loss in 50-taxon, the ATE rate of the former was much higher than that of the latter two in 100-taxon data sets (Figure 8).

We also examined the accuracy of species tree estimates by Only-dup, Dup-loss, and SPRS when gene tree simulations include events that these methods do not assume to be the source of discordance (e.g., *dl* and *both* for SPRS, *lgt* and *both* for Only-dup and Dup-loss). While SPRS could not be tested on *dl* and *both* because they included mul-trees, Only-dup and Dup-loss had high ATE rate for *lgt* and *both* (Figure 8). In general, Only-dup's estimate had much higher ATE rate compared to Dup-loss in the presence of LGT events; the ATE rate of MulRF was lowest among all the methods (Figure 8).

Biological data set

We also tested the performance of the MulRF method on a gene tree data set from 22 plant species. These species were chosen because they are phylogenetically diverse, and they all have fully sequenced and annotated genome sequences. This makes it possible to obtain a large number of gene trees with potentially no missing sequences. Furthermore, there is much support for the relationships among most of these species (e.g., [45]), and therefore, it provides an empirical system on which we can evaluate the performance of MulRF. We obtained nucleotide alignments from gene families that had been generated from genome sequences with OrthMCL [46] and aligned with MAFFT [47]. We selected the gene family alignments that contained sequences from at least 20 of the 22 species and had a maximum of 50 gene sequences. This was a total of 1910 gene alignments. We estimated maximum likelihood trees for all of the genes using GTRCAT model in RAxML [43]. The unrooted gene trees were used as input for the MulRF heuristic. The Only-dup and Dup-loss methods require rooted input trees. Thus, we rooted all of the gene trees on the longest branch using Newick utilities [48]. This is similar to mid-point rooting, and in our experience it often provides a good starting point for input gene trees in GTP analyses. The rooted gene trees were used as input for GTP analysis using Only-dup [13] and Dup-loss [14]. Only-dup and Dup-loss were executed with the default SPR search, including a fast leaf-adding heuristic for initial species tree construction, and searching for an optimal root by re-rooting the gene trees after each SPR search (e.g., [13]; [17]). We could not run SPRS on this data set because it contains mul-trees.

The MulRF heuristic completed in 4 minutes and 4 seconds on a Mac laptop with a 2.26 GHz Intel processor and 4GB RAM. The resulting species tree is largely consistent with the most recent phylogenetic analyses (Figure 9; e.g., [45]). *Amborella* is sister to the other angiosperms and monocots and eudicots form clades. Within the eudicots there is a core-eudicot clade, and within the core-eudicots the rosid clade is sister to the asterid clade. The malvids are sister to the fabids within the rosids. Interestingly, *Populus* groups with the malvids, consistent with recent analyses of nuclear and mitochondrial, but not chloroplast, data (e.g., [49]; [17]). There are two minor differences from the generally accepted reltionships: Phoenix should be sister to Musa + Poaceae rather than sister to Musa, and Aquilegia should be sister to the other eudicots rather than Nelumbo. The ATE for the MulRF tree is 0.11. Thus, it appears that MulRF can quickly estimate a nearly accurate species trees from large-scale plant genomic data sets. The Only-dup tree heuristic completed in 7 seconds, and if we unroot the result, it is identical to the MulRF tree. The Dup-loss tree, which completed in 7 seconds, had a less accepted topology, placing *Amborella* sister to the monocots instead of sister to other angiosperms and

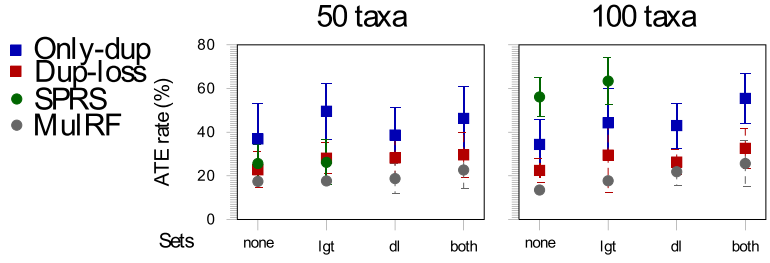

Figure 8 Experimental result of simulated data. Average topological error (means with standard error bars) for species tree constructed by Only-dup, Dup-loss, SPRS, and MulRF method, for all four model conditions.

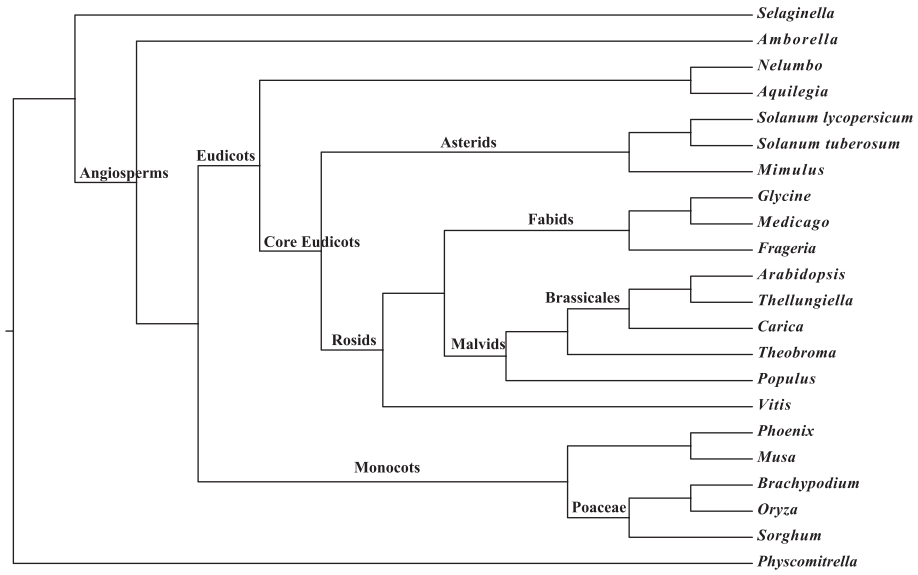

Figure 9 Experimental result of biological data. The MulRF species tree of plant gene tree data set.

Vitis sister to the asterids rather than with the rosids. The ATE for the Dup-loss tree is 0.21.

Conclusion

We presented the new MulRF method for inferring species tree from incongruent gene trees that is based on a generalized form of the RF distance. Unlike most previous phylogenetic methods using gene trees, our approach is based on a generic tree distance measure that is not linked to any specific biological processes. As a result, it is intuitively appealing for analyses of genomic data sets, in which many processes such as deep coalescence, recombination, gene duplications and losses, and LGT, as well as phylogenetic error likely contribute to gene tree discord. In simulation experiments, the MulRF method estimated species trees more accurately than several GTP methods, and it appears to be relatively robust to the effects of phylogenetic error, gene duplication and loss, and LGT. In addition, the MulRF method is fast, estimating 100-taxon species trees from hundreds of gene trees in under two minutes and a plant data set with 22 taxa and nearly 2000 gene trees in just over 4 minutes.

Our simulation experiments greatly simplify the true processes of genomic evolution. We focused only on processes that reflect the objectives of the GTP methods, and we emphasized on duplication and loss, because that especially relevant to the evolution of multi-copy gene trees. Still, even in these conditions in which we might expect GTP to perform well, we find that MulRF obtains more accurate results than GTP in most instances. This does not mean that MulRF will always outperform GTP, but we suggest that MulRF can quickly provide an interesting alternate perspective on species tree inference. More tests are needed to characterize the performance of MulRF methods under different evolutionary scenarios.

Another future direction will be to incorporate estimates of gene tree uncertainty into the supertree analysis by weighing the splits differently when computing the RF distance. Also, the effectiveness of the MulRF method in inferring species trees from multi-copy gene trees suggests that other tree distance measures can be used in the same context. A natural candidate for study is the quartet distance. Future work should also evaluate the suitability of different distance metrics in estimating species trees under different error models and evolutionary scenarios.

Additional file

Additional file 1: Inferring species trees from incongruent multi-copy gene trees using the Robinson-Foulds distance. A pdf file containing the proof of Theorem 1.

Competing interests

The authors declare that they have no competing interests.

Authors' contributions

RC designed and implemented the algorithm, developed the NP-completeness proof, performed simulation experiments, and wrote major parts of the paper. JGB conducted experiments on biological data set and contributed to the writing of the paper. DFB supervised the project and contributed to the writing of the manuscript. All authors read and approved the final manuscript.

Acknowledgements

This work was supported in part by NSF grants DEB-0829674 (to D.F.-B), CCF-1048217 (to J.G.B), and DEB-1208428 (to J.G.B.). We thank the reviewers for carefully reading the entire manuscript and offering many useful suggestions.

Received: 28 May 2013 Accepted: 8 October 2013
Published: 1 November 2013

References

1. Maddison WP: **Gene trees in species trees.** *Syst Biol* 1997, **46:**523–536.
2. Avise J, Shapira J, Daniel S, Aquadro C, Lansman R: **Mitochondrial DNA differentiation during the speciation process in Peromyscus.** *Mol Biol Evol* 1983, **1:**38–56.
3. Doyle J: **Gene trees and species trees: molecular systematics as one-character taxonomy.** *Syst Bot* 1993, **17:**144–163.
4. Goodman M, Czelusniak J, Moore GW, Romero-Herrera AE, Matsuda G: **Fitting the gene lineage into its species lineage. A parsimony strategy illustrated by cladograms constructed from globin sequences.** *Syst Zool* 1979, **28:**132–163.
5. Maddison W: **Molecular approaches and the growth of phylogenetic biolog.** In *Molecular Zoology: Advances, Strategies and Protocols*. Edited by Ferraris JD, Palumbi SR. New York: Wiley-Liss; 1996:47–63.
6. Pamilo P, Nei M: **Relationships between gene trees and species trees.** *Mol Biol Evol* 1988, **5:**568–583.
7. Kubatko LS, Carstens BC, Knowles LL: **STEM: species tree estimation using maximum likelihood for gene trees under coalescence.** *Bioinformatics* 2009, **25**(7):971–973.
8. Boussau B, Szöllősi GJ, Duret L, Gouy M, Tannier E, Daubin V: **Genome-scale coestimation of species and gene trees.** *Genome Res* 2012, **23:**323–330.
9. Ané C, Larget B, Baum DA, Smith SD, Rokas A: **Bayesian estimation of concordance among gene trees.** *Mol Biol Evol* 2007, **24**(7):1575.
10. Liu L, Pearl DK: **Species trees from gene trees: reconstructing Bayesian posterior distributions of a species phylogeny using estimated gene tree distributions.** *Syst Biol* 2007, **56**(3):504–514.
11. Drummond AJ, Rambaut A: **BEAST: Bayesian evolutionary analysis by sampling trees.** *BMC Evol Biol* 2007, **7:**214.
12. Page RDM: **GeneTree: comparing gene and species phylogenies using reconciled trees.** *Bioinformatics* 1998, **14**(9):819–820.
13. Wehe A, Bansal MS, Burleigh JG, Eulenstein O: **DupTree: a program for large-scale phylogenetic analyses using gene tree parsimony.** *Bioinformatics* 2008, **24:**13.
14. Bansal MS, Burleigh JG, Eulenstein O: **Efficient genome-scale phylogenetic analysis under the duplication-loss and deep coalescence cost models.** *BMC Bioinformatics* 2010, **11**(Suppl 1):S42.
15. Yu Y, Warnow T, Nakhleh L: **Algorithms for MDC-based multi-locus phylogeny inference.** In *RECOMB*. Heidelberg: Springer-Verlag Berlin; 2011:531–545.
16. Whidden C, Zeh N, Beiko R: **SPRSupertrees. Version 1.1.0.** 2012. [http://kiwi.cs.dal.ca/Software/SPRSupertrees]
17. Burleigh JG, Bansal MS, Eulenstein O, Hartmann S, Wehe A, Vision TJ: **Genome-scale phylogenetics: inferring the plant tree of life from 18,896 discordant gene trees.** *Syst Biol* 2011, **60**(2):117–125.
18. Huang H, Knowles LL: **What is the danger of the anomaly zone for empirical phylogenetics?** *Syst Biol* 2009, **58:**527–536.
19. Sanderson MJ, McMahon MM: **Inferring angiosperm phylogeny from EST data with widespread gene duplication.** *BMC Evol Biol* 2007, **7**(suppl 1):S3. [http://www.biomedcentral.com/1471-2148/7/S1/S3]
20. Than C, Rosenberg N: **Consistency properties of species tree inference by minimizing deep coalescences.** *J Comput Biol* 2011, **18:**1–15.
21. Cui Y, Jansson J, Sung WK: **Algorithms for building consensus MUL-trees.** In *International Symposium on Algorithms and Computation (ISAAC'2011)*, LNCS 7074. Heidelberg: Springer-Verlag Berlin; 2011:744–753.
22. Cui Y, Jansson J, Sung W: **Polynomial-time algorithms for building a consensus MUL-tree.** *J Comput Biol* 2012, **19:**1073–1088.
23. Huber KT, Lott M, Moulton V, Spillner A: **The complexity of deriving multi-labeled trees from bipartitions.** *J Comput Biol* 2008, **15:**639–651.
24. Huber K, Moulton V, Spillner A: **Computing a consensus of multilabeled trees.** In *Proceedings of the 14th Workshop on Algorithm Engineering and Experiments (ALENEX 2012)*; 2012:84–92.
25. Huber KT, Spillner A, Suchecki R, Moulton V: **Metrics on multilabeled trees: interrelationships and diameter bounds.** *IEEE/ACM Trans Comput Biol Bioinform* 2011, **8:**1029–1040.
26. Guillemot S, Jansson J, Sung WK: **Computing a smallest multilabeled phylogenetic tree from rooted triplets.** *IEEE/ACM Trans Comput Biol Bioinform* 2011, **8:**1141–1147.
27. Bogdanowicz D, Giaro K: **Matching split distance for unrooted binary phylogenetic trees.** *IEEE/ACM Trans Comput Biol Bioinform* 2012, **9:**150–160.
28. Lin Y, Rajan V, Moret BM: **A metric for phylogenetic trees based on matching.** *IEEE/ACM Trans Comput Biol Bioinform* 2012, **9:**1014–1022.
29. Bansal MS, Burleigh JG, Eulenstein O, Fernández-Baca D: **Robinson-Foulds supertrees.** *Algorithms Mol Biol* 2010, **5:**18.
30. Chaudhary R, Burleigh JG, Fernández-Baca D: **Fast local search for unrooted Robinson-Foulds supertrees.** *IEEE/ACM Trans Comput Biol Bioinform* 2012, **9:**1004–1013.
31. Steel M, Rodrigo A: **Maximum likelihood supertrees.** *Syst Biol* 2008, **57:**2.
32. Semple C, Steel M: *Phylogenetics*. New York: Oxford University Press Inc; 2003.
33. Robinson DF, Foulds LR: **Comparison of phylogenetic trees.** *Math Biosci* 1981, **53:**131–147.
34. Ganapathy G, Goodson B, Jansen R, Le H, Ramachandran V, Warnow T: **Pattern identification in biogeography.** *IEEE/ACM Trans Comput Biol Bioinform* 2006, **3:**334–346.
35. McMorris FR, Steel MA: **The complexity of the median procedure for binary trees.** In *Proceedings of the International Federation of Classification Societies*. Heidelberg: Springer-Verlag Berlin; 1994.
36. Allen BL, Steel M: **Subtree transfer operations and their induced metrics on evolutionary trees.** *Ann Combinatorics* 2001, **5:**1–15.
37. Bender MA, Farach-Colton M: **The LCA Problem Revisited.** In *LATIN, Volume 1776 of Lecture Notes in Computer Science*. Edited by Gonnet GH, Panario D, Viola A. Heidelberg: Springer-Verlag Berlin; 2000:88–94.
38. Maddison WP, Maddison D: **Mesquite: a modular system for evolutionary analysis. Version 2.6.** 2009. [http://mesquiteproject.org].
39. Arvestad L, Berglund A-C, Lagergren J, Sennblad B: **Bayesian gene/species tree reconciliation and orthology analysis using MCMC.** *Bioinformatics* 2003, **19**(Suppl 1):i7–i15.
40. Rambaut A, Grassly NC: **Seq-Gen: An application for the Monte-Carlo simulation of DNA sequence evolution along phylogenetic trees.** *Copmput Appl Biosci* 1997, **13:**235–238.
41. Ganapathy G: **Algorithms and heuristics for combinatorial optimization in phylogeny.** PhD thesis, University of Texas at Austin 2006.
42. Swenson MS, Barbançon F, Warnow T, Linder CR: **A simulation study comparing supertree and combined analysis methods using SMIDGen.** *Algorithms Mol Biol* 2010, **5:**8.
43. Stamatakis A: **RAxML-VI-HPC: Maximum likelihood- based phylogenetic analyses with thousands of taxa and mixed models.** *Bioinformatics* 2006, **22:**2688–2690.
44. Felsenstein J: **Retree software.** 1993. [http://evolution.genetics.washington.edu/phylip/doc/retree.html]
45. Soltis DE, Smith SA, Cellinese N, Wurdack KJ, Tank DC, Brockington SF, Refulio-Rodriguez NF, Walker JB, Moore MJ, Carlsward BS, Bell CD, Latvis M, Crawley S, Black C, Diouf D, Xi Z, Rushworth CA, Gitzendanner MA, Sytsma KJ, Qiu YL, Hilu KW, Davis CC, Sanderson MJ, Beaman RS, Olmstead RG, Judd WS, Donoghue MJ, Soltis PS: **Angiosperm phylogeny: 17 genes, 640 taxa.** *Am J Bot* 2011, **98:**704–730.
46. Chen F, Mackey AJ, Stoeckert CJ Jr, Roos DS: **OrthoMCL-DB: querying a comprehensive multi-species collection of ortholog groups.** *Nucleic Acids Res* 2006, **34:**D363–D368.
47. Katoh K, ichi Kuma, K, Toh H, Miyata T: **MAFFT version 5: improvement in accuracy of multiple sequence alignment.** *Nucleic Acids Res* 2005, **33:**511–518.
48. Junier T, Zdobnov EM: **The Newick utilities: high-throughput phylogenetic tree processing in the Unix shell.** *Bioinformatics* 2010, **26:**1669–1670.
49. Qiu YL, Li L, Wang B, Xue JY, Hendry TA, Li RQ, Brown JW, Liu Y, Hudson GT, Chen ZD: **Angiosperm phylogeny inferred from sequences of four mitochondrial genes.** *J Syst Evol* 2010, **48:**391–425.

doi:10.1186/1748-7188-8-28
Cite this article as: Chaudhary *et al.*: Inferring species trees from incongruent multi-copy gene trees using the Robinson-Foulds distance. *Algorithms for Molecular Biology* 2013 **8**:28.