ALGORITHMIC STRATEGIES FOR ESTIMATING THE AMOUNT OF RETICULATION FROM A COLLECTION OF GENE TREES

H. J. Park and G. Jin and L. Nakhleh*

Department of Computer Science, Rice University,
6100 Main Street, Houston, TX 77005, USA
**Email: nakhleh@cs.rice.edu*

Phylogenetic networks have emerged as a unifying evolutionary model of both vertical and horizontal inheritance. A major approach for reconstructing such networks is to reconcile gene trees that are reconstructed from various genomic regions. The Subtree Prune and Regraft (SPR) operation has been used to obtain lower bound estimates of the number of reticulation events from a pair of trees. However, more than two trees are available in general and, to date, no work exists on estimating the amount of reticulation by the SPR operation from a collection, not only a pair, of trees.

In this paper we address this problem, and propose two algorithmic strategies for heuristically solving it. The first is based on a simple, yet novel, observation on the binomial distribution of pairwise distances of trees inside a network. The second is based on the aggregation of solutions from pairwise computations. We have implemented both approaches and studied their performance in extensive simulations. The methods produce good results in general in terms of estimating the minimum number of reticulation events required to reconcile a set of trees. In addition, we identify conditions under which the methods do not work as well, in an attempt to help in the development of new methods in this area.

1. INTRODUCTION

When the evolution of a group of species involves *reticulate*, or non-treelike, evolutionary events, such as horizontal gene transfer or hybridization, the evolutionary history of the species is best modeled by a *phylogenetic network*—a rooted, directed, acyclic graph, leaf-labeled by a set of taxa. However, even though the phylogenomic history of the organisms in this case takes the shape of a network, the species' genomes can be partitioned into regions each of which has a treelike evolutionary history. This observation has been the basis for the tree-based approach to phylogenetic network reconstruction: (1) infer the evolutionary trees for the different genomic regions (ideally, those genomic regions are recombination-free), and (2) combine the set of trees into a phylogenetic network that satisfies some criterion. Extensive surveys of phylogenetic network models, issues, and reconstruction methods, have recently appeared in the literature [1–7].

Several methods have been developed for inferring a lower bound on the number of reticulation events by identifying the minimum number of *subtree prune and regraft*, or SPR, operations required to transform one tree into the other [8]. An SPR operation applied to tree T cuts, or prunes, a subtree t of T, yielding a tree T', and attaches, or regrafts, it from its root to another branch in T' [9]. Underestimation issues notwithstanding [10, 11, 7], the SPR-based approach has been heavily used as a proxy for inferring (lower bound on) the number of reticulation events, as well as their placement in the evolutionary history. The problem of computing the SPR distance between two trees has been shown to be NP-hard as well as fixed-parameter tractable [12]. Examples of exact algorithms and heuristics for reconciling trees via SPR operations include the exact algorithm of Bordewich and Semple [12], the exact algorithm of Wu [13], LatTrans [14], RIATA-HGT [15, 16], EEEP [17], HorizStory [18], and the method of Goloboff as implemented in the TNT software package [19]. *However, one salient feature of all these methods is that they only apply to a pair, but not a larger set, of trees.*

With the availability of whole-genome data from an increasingly large number of organisms, particularly prokaryotic ones, evolutionary studies are

[*] Corresponding author.

[a] In this context, the term "gene tree" applies to an evolutionary tree of any non-recombining genomic region; i.e., it is not limited to trees on (protein-coding) gene regions.

faced with a large number of *gene trees*[a] in a given study. Therefore, it is imperative to develop computational techniques that simultaneously analyze a large number of trees, and combine them into networks. Clearly, the problem is NP-hard when the SPR distance is used, since it is NP-hard for a pair of trees. Huson and Rupp [20] proposed a method for summarizing a collection of gene trees using *cluster networks*, which differ from the phylogenetic network model we address here. More Recently, Beiko and Ragan [21] discussed aggregating inferred HGT events from pairwise tree comparisons, and discussed three strategies for this task; yet, they did not implement the strategies, nor did they study their performance. In this paper, we address the problem of inferring a phylogenetic network with the minimum number of reticulation events that reconciles a collection of gene trees using the SPR operation. We present two heuristic algorithms, one that is based on the observation of a binomial distribution of the pairwise distances of a collection of trees contained in a network, and the second is based on agglomerating pairwise solutions to obtain a global, hopefully minimal, solution for all trees. We have implemented both algorithms and studied their performance, in terms of the number of reticulations they infer from a set of trees, on a large number of simulated data sets. Results indicate very good performance of the methods in general, and highlight conditions under which the methods' performance is not as good. The latter issue is particularly important, since it may help develop more accurate methods for this problem.

The rest of the paper is organized as follow. In Section 2 we give an explicit definition of the phylogenetic network model that we use in this paper, and its relationship to trees. This is very important, since the term 'phylogenetic network' has been used in different contexts to mean different things as well as to have different properties [7]. In Sections 3 and 4, we describe two algorithmic strategies for estimating the minimum number of reticulations based on the distribution of pairwise SPR distances and the agglomeration of pairwise SPR moves, respectively. In Section 5 we demonstrate the performance of both algorithmic strategies on a large number of simulated data. We conclude in Section 6 with final remarks and some directions for future research.

2. BACKGROUND

In this paper, we focus on rooted, binary trees and networks.

Definition 2.1. A phylogenetic *$\mathcal{X}$-network*, or *$\mathcal{X}$-network*, N is an ordered pair (G, f), where

(1) G = (V, E) is a directed, acyclic graph (DAG) with V $= \{r\} \cup V_L \cup V_T \cup V_N$, where
   - (a) $indeg(r) = 0$ (r is the *root* of N);
   - (b) $\forall v \in V_L, indeg(v) = 1$ and $outdeg(v) = 0$ (V_L are the leaves of N);
   - (c) $\forall v \in V_T, indeg(v) = 1$ and $outdeg(v) \geq 2$ (V_T are the tree-nodes of N); and,
   - (d) $\forall v \in V_N, indeg(v) = 2$ and $outdeg(v) \geq 1$ (V_N are the network-nodes of N),
   - (e) and $E \subseteq V \times V$ are the network's edges. (we distinguish between network-edges, edges whose heads are network-nodes, and tree-edges, edges whose heads are tree-nodes or leaves.)

(2) f: $V_L \to \mathcal{X}$ is a bijection function from V_L to $\mathcal{X}$.

A phylogenetic $\mathcal{X}$-tree is an $\mathcal{X}$-network in which $V_N = \emptyset$. While a network N represents the evolution of a set of genomes, these genomes can be partitioned into (non-recombining) regions $R_1, R_2, \ldots, R_k$, each of which has a treelike evolutionary history T_i. In other words, the set $\mathcal{T} = \{T_1, \ldots, T_k\}$ is a subset of the set of all trees *induced* by the network N. More formally, $\mathcal{T} \subseteq \mathcal{T}(N)$, where $\mathcal{T}(N)$ is the set of *all* trees obtained as follows from N: (1) for each node of in-degree 2 remove one of the two incoming edges and (2) for each node u of in-degree and out-degree 1, remove u along with its incident edges, and add a new edge to connect u's parent to u's child (this step is repeated until no such nodes u remain).

Given an $\mathcal{X}$-network N, it is straightforward to compute the set $\mathcal{T}(N)$, though this computation may be expensive, since $|\mathcal{T}(N)| = O(2^{|V_N|})$. The more relevant problem in the context of inferring phy-

[b]It is highly unlikely for a biological data set to exhibit all trees induced by the network; in practice, the set of trees exhibited by the different genomic regions is a small subset of all possible trees induced by the network.

logenetic relationships is that of estimating an $\mathcal{X}$-network from a subset[b] of its induced trees, since this amounts to inferring the (reticulate) evolutionary history of a set of organisms.

Problem 2.1. *Given a set of $\mathcal{X}$-trees $\mathcal{T} = \{T_1, T_2, \ldots, T_k\}$, each modeling the evolutionary history of a genomic region, we seek the $\mathcal{X}$-network that models the evolutionary history of the genomes.*[c]

Obviously, if all trees in $\mathcal{T}$ are identical, the problem is trivial since N would be the tree in $\mathcal{T}$. Otherwise, the problem is hard. Given that there is a very large number of $\mathcal{X}$-networks N such that $\mathcal{T} \subseteq \mathcal{T}(N)$, the main issue in this domain is to define a criterion Φ and seek the $\mathcal{X}$-network (or, set of $\mathcal{X}$-networks) that is optimal Φ, given the set $\mathcal{T}$ of trees. A natural parsimony criterion to define is to minimize the number of network-nodes in N. In other words, we seek the network (or set of networks) N such that (1) $\mathcal{T} \subseteq \mathcal{T}(N)$, and (2) N has the minimum number of network-nodes among all $\mathcal{X}$-networks satisfying (1). While the "true" phylogenetic network may not necessarily be a parsimonious one, this criterion yields plausible networks in many realistic cases (although it is easy to show examples of cases in which this criterion results in networks with numbers of network-nodes that are arbitrarily smaller than the true number [7]). In particular, this criterion can be viewed as a way to estimate a lower bound on the amount of reticulation in the data. When $\mathcal{T} = \{T_1, T_2\}$, a solution to the problem is to compute the *SPR distance* [9] between the two trees, denoted by $d_{SPR}(T_1, T_2)$, and take it as the estimate of the number of network-nodes in the $\mathcal{X}$-network N that induced both trees in $\mathcal{T}$. An SPR (Subtree Prune and Regraft) operation applied to a tree T_1 is defined by *pruning* a subtree t in T_1, and *regrafting* t from its root to another branch in T_1 (that is not in t). Given two $\mathcal{X}$-trees T_1 and T_2, T_1 can be transformed into T_2 by a sequence of SPR moves, and the length of the shortest such sequence is defined as the SPR distance between the two trees.

In this paper we address the problem of estimating an $\mathcal{X}$-network, with the minimum number of network-nodes, that induces a given set of trees $\mathcal{T}$. This problem is NP-hard, given that is NP-hard for a pair of trees [12]. As we show below, our investigation of simulated data sets indicates that, in practice, one factor that may affect the hardness of the problem is the *redundancy* in the network, which we define as follows.

Definition 2.2. The redundancy of an $\mathcal{X}$-network N with set V_N of network-nodes is $\varepsilon_N = (2^{|V_N|} - |\mathcal{T}(N)|)/2^{|V_N|}$.

Figure 1 illustrates the concept of redundancy.

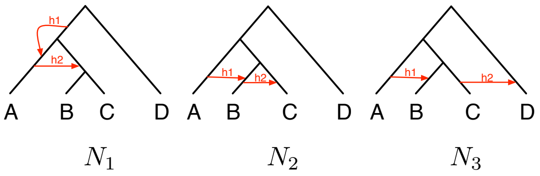


Fig. 1. Three $\mathcal{X}$-networks, each with two network-nodes, yet with varying degrees of redundancy. Here, $\mathcal{T}(N_1) = \{T_1\}$, $\mathcal{T}(N_2) = \{T_1, T_2\}$, and $\mathcal{T}(N_3) = \{T_1, T_2, T_3, T_4\}$, where $T_1 = ((A,(B,C)),D)$, $T_2 = (((A,B),C),D)$, $T_3 = (A,(B,(C,D)))$, and $T_4 = ((A,B),(C,D))$. Consequently, we have $\varepsilon_{N_1} = (4-1)/4 = 0.75$, $\varepsilon_{N_2} = (4-2)/4 = 0.50$, and $\varepsilon_{N_3} = (4-4)/4 = 0$.

In a non-redundant $\mathcal{X}$-network N ($\varepsilon_N = 0$), each tree in $\mathcal{T}(N)$ is uniquely induced by the network, whereas in a redundant network ($\varepsilon_N > 0$), some trees may be induced in multiple ways. An upper bound on ε_N for an $\mathcal{X}$-network with h network-nodes is $1 - 1/2^h$, in which case the network induces a single tree and, considering topology alone, none of the reticulation events may be detectable.

3. FITTING A BINOMIAL DISTRIBUTION OF PAIRWISE DISTANCES

Let $V_N = \{v_1, \ldots, v_h\}$ be the set of all network-nodes in an $\mathcal{X}$-network N, and for each two edges incoming into a node $v_i \in V_N$, let one be labeled l (for *left*) and the other be labeled r (for *right*). Further, let $T \in \mathcal{T}(N)$ be a tree induced by the network. A *displaying vector* of T, denoted by $d(T)$ is an element of $\{l, r\}^h$, where $d(T)[i]$ denotes the label of the edge incoming into v_i that was retained to induce the tree

[c] It is important to note that while we focus on collections of trees that have the same leaf labels, in practice gene trees may differ in their leaf labels (e.g., due to sampling, gene duplication, gene loss, etc.).

T. We have the following two lemmas and ensuing theorem.

Lemma 3.1. *Let N be an $\mathcal{X}$-network. Then, $d(T)$ is unique for every tree $T \in \mathcal{T}(N)$ iff $\varepsilon_N = 0$.*

Lemma 3.2. *Let $\mathcal{D} = \{l, r\}^h$. Then $|\{\{d_1, d_2\} : d_1, d_2 \in \mathcal{D},\ HD(d_1, d_2) = q\}| = \binom{h}{q}2^{h-1}$, where $HD(d_1, d_2)$ denotes the Hamming distance between the two binary vectors d_1 and d_2.*

Theorem 3.1. *Let N be an $\mathcal{X}$-network with h network-nodes, and assume $d_{SPR}(T_1, T_2) = HD(d(T_1), d(T_2))$ for every $T_1, T_2 \in \mathcal{T}(N)$. If $\varepsilon_N = 0$ then $|\{\{T_1, T_2\} : T_1, T_2 \in \mathcal{T}(N),\ d_{SPR}(T_1, T_2) = q\}| = \binom{h}{q}2^{h-1}$.*

Theorem 3.1 implies that when there is no redundancy in the network, and given that we do not know the actual displaying vectors of the trees, we can use the SPR distance as a proxy to the Hamming distance of the displaying vector, and expect a binomial distribution of the pairwise distances. This, in turn, naturally gives rise to the following approach for estimating the minimum number of reticulations required in a phylogenetic network to reconcile a set $\mathcal{T}$ of trees:

(1) Compute all pairwise SPR distances over the set $\mathcal{T}$ of trees, and let Q be the distribution of these distances.
(2) Denoting by P_m the distribution $\binom{m}{q}2^{m-1}$ for $1 \leq q \leq m$, find the value m that minimizes $KL(Q|P_m)$, where KL is the Kullback-Leibler distance [22] $KL(g|f) = \sum_q f(q) \ln \frac{f(q)}{g(q)}$.

The way we compute the value of m in Step (2) in the above procedure is by starting from

$$m = \max\{\lceil \log_2 |\mathcal{T}| \rceil, \max_{T_1, T_2 \in \mathcal{T}} d_{SPR}(T_1, T_2)\} \qquad (1)$$

and incrementing m as long as $KL(Q|P_m)$ decreases. The rationale behind Equation (1) is that the $\log_2$ of the number of trees in the given set is a lower bound on the number of reticulations, and so is the maximum pairwise SPR distance over all trees in the set.

Obviously, the conditions of Theorem 3.1 may not hold in practice. In particular, it may be that some or all of the following issues arise when analyzing a data set:

(1) It may be that for some pairs of trees $T_1, T_2 \in \mathcal{T}(N)$, $d_{SPR}(T_1, T_2) < HD(d(T_1), d(T_2))$. In this case, the distribution of the pairwise distances may be skewed to the left. A potential alternative for considering the minimum number of SPR moves is to take a stochastic approach that simulates random walks, using SPR moves, in the tree space [23].
(2) The (unknown) network N may have $\varepsilon_N > 0$. Here, the frequencies of some pairwise distances may be lower than the true frequencies (which are the ones based on P_m).
(3) The given set of trees $\mathcal{T}$ does not contain all trees induced by the (unknown) network N. Here, not enough data points may be available for reliably estimating the true distribution Q.

Nevertheless, we show below, through extensive simulations, that this heuristic provides good estimates of the number of network-nodes required for a network to reconcile a given set of trees.

4. COMBINING PAIRWISE SOLUTIONS

While the approach in the previous section is aimed at estimating only the minimum number of reticulations needed in a phylogenetic network to reconcile a set of trees, the approach we present here is aimed at estimating minimal sets of actual SPR moves (obviously, the sizes of such sets can be taken as estimates of the amount of reticulation). The general outline of the method we propose here for estimating a set of SPR moves to reconcile a set of trees $\mathcal{T}$ is simple (similar to the *greedy approach* for aggregating inferred HGT events in [21]):

(1) For each pair of trees in $\mathcal{T}$, identify a minimal set of SPR moves that reconcile them.
(2) Combine the set of solutions identified in Step (1).

There are two main issues that need to be addressed for this approach to work in practice. First, for a given pair of trees, there may be multiple minimal sets of SPR moves that reconcile them [24]. In this case, we need the pairwise SPR computation to return all, or a large number, of these minimal solutions. We make use of the modified version of

RIATA-HGT [15, 25], as implemented in PhyloNet [16], to compute multiple minimal solutions. The second issue is two-fold: (a) Given a set of minimal sets of SPR moves for each pair of trees, how do we find a global minimal set of SPR moves that covers at least one minimal set for each pair? (b) Once the (global) minimal set is computed, how do we obtain a network from it?

In the case of the horizontal gene transfer detection problem, usually a species tree ST is given, in addition to the set of trees $\mathcal{T}$. In this case, the pairwise computations should be conducted only between ST and every tree in $\mathcal{T}$, but not between pairs of trees in $\mathcal{T}$. Then, the global set of SPR moves computed by the procedure above is posited on the tree ST. In the case where no "backbone" tree, such as ST, is given, we propose to use each of the k trees in $\mathcal{T}$ as a backbone tree against each all SPR computations are conducted, and choose the tree in $\mathcal{T}$ that results in the smallest set of SPR moves.

We use this idea in the heuristic **ComputeSPRsMultiGenes** below. Let ST be an (species) $\mathcal{X}$-tree and $\mathcal{T} = \{T_1, \ldots, T_k\}$ be a collection of (gene) $\mathcal{X}$-trees. Further, let Z be the set of all possible SPR moves that can be defined on ST (the cardinality of Z is quadratic in the number of leaves in ST [9]). For each tree $T_i \in \mathcal{T}$, let $SPR(ST, T_i) = \{S_i^1, \ldots, S_i^{w_i}\}$ be the set of minimal sets of SPR moves that transform ST into T_i. Our task is to find a minimal set $z \subseteq Z$ such that for every $1 \leq i \leq k$, there exists $1 \leq \ell_i \leq w_i$ such that $S_i^{\ell_i} \subseteq z$. In other words, we seek a minimal set z of SPR moves that cover at least one minimal "solution" for each gene tree. Clearly, each tree in $\mathcal{T}$ can be obtained by applying a subset (or all) of the SPR moves in z to ST. This is a hard problem, and we solve it heuristically, as described in the following algorithm.

```
ALGORITHM ComputeSPRsMultiGenes
1. For each gene tree $T_i \in \mathcal{T}$
    1.1. initialize count: $c(r) = 0$ for every SPR move $r$ in $Z$;
    1.2. for each gene tree $T_j \in \mathcal{T}$ and $T_j \neq T_i$
        1.2(a). compute $SPR(T_i, T_j)$;
    1.3. for each SPR move $r$, compute count $c(r) = |\{j | r \in$ solution $s$ and $s \in SPR(T_i, T_j)\}|$;
    1.4. for each gene tree $T_j \in \mathcal{T}$and $T_j \neq T_i$
        1.4(a). for each solution $s \in SPR(T_i, T_j)$, compute count $c(s) = \sum_k \{c(r_k) | r_k \in s\}$;
        1.4(b). choose a solution $s$, $\hat{SPR}(T_i, T_j) = \{s | c(s) >= c(s')$ for all $s' \neq s, s' \in SPR(T_i, T_j)\}$;
    1.5. compute the union $R_i = \bigcup_{T_j \in \mathcal{T}, T_j \neq T_i} \{s | s \in \hat{SPR}(T_i, T_j)\}$;
2. choose $\mathcal{R} = R_l$ such that $|R_l| = min_i(|R_i| | 1 \leq i \leq k)$ along with the corresponding tree $T_l \in \mathcal{T}$.
```

5. EXPERIMENTAL EVALUATION

5.1. Experimental Setup

To simulate phylogenetic networks, we used two tools: the PHYL-O-GEN tool [26] for generating random "species trees" under the birth-death model, and the tool of Galtier [27] to simulate horizontal gene transfer events (HGTs) between pairs of branches of the species trees. Since Galtier's tool adds a random number of HGTs and does not report the number, or placement, of those events, we modified the tool so that the output includes the actual HGT edges that it adds. The direct parameters in our experiments are the number of taxa and the number of HGTs simulated. For the number of taxa (leaves in the trees and networks), we used 10, 20, 30, and 50, and for the number of HGTs, we used 5 (for trees with 10, 20, and 30 leaves) and 10 (for trees with 50 leaves). For each number of taxa, we generated 10 trees with that number of leaves. Each such (species) tree was used as an input to Galtier's tool to simulate HGTs and create networks. For each (species) tree and specific number of HGTs, 12 networks were generated. In total, for each of the combinations (10 taxa, 5 HGTs), (20 taxa, 5 HGTs), (30 taxa, 5 HGTs), and (50 taxa, 10 HGTs), we generated 120 networks, for a total of 480 networks (in the case of 30 taxa, pairwise SPR distances were overestimated in certain cases, which we removed from the analysis, to control for problems with pairwise distance estimation. As a result, the number of 30-taxon networks was 80, instead of 120). For each network N, we computed, using PhyloNet [16], the set of its induced trees $\mathcal{T}(N)$, and sampled from this set (so as to simulate phylogenomic trees) collections of gene trees, and gave those as inputs to our methods. In other words, each method was run on a collection of trees generated from network. To obtain statistically significant results, we repeated the sampling process 30 times for each sample size, and plotted the averages.

An explanation on what we measure is in order. When either of the two methods is run on a collection $\mathcal{T} = \{T_1, \ldots, T_k\} \subseteq \mathcal{T}(N)$ induced by network N, we record the number of reticulations that the method computed; we call this number the *detected* number of reticulations. Now, if network N was generated with 5 or 10 HGTs, this does not necessarily mean that the collection $\mathcal{T}$ of trees will have all trees to allow for detecting 5 or 10 HGTs, respectively. For example, consider the collection $\mathcal{T}$ that has only trees whose (pairwise) SPR distance is 1. In this case, the number of detectable HGTs is 1, and not 5 (or 10). Therefore, for each such collection $\mathcal{T}$, we compute (exhaustively) the smallest subset of HGTs in N that can reconcile all trees in $\mathcal{T}$; we call this number the *detectable* number of reticulations (notice that this is not necessarily the smallest number of reticulations needed to reconcile all trees in $\mathcal{T}$; computing this number would be prohibitive). The accuracy of method is considered better as the difference between the detectable and detected numbers of reticulations becomes smaller.

We ran both methods on all data sets. In the next section, we refer to the method that fits the pairwise distances to a binomial distribution **Method M1** (see Section 3), and to the method based on the union of pairwise SPR move sets **Method M2** (see Section 4).

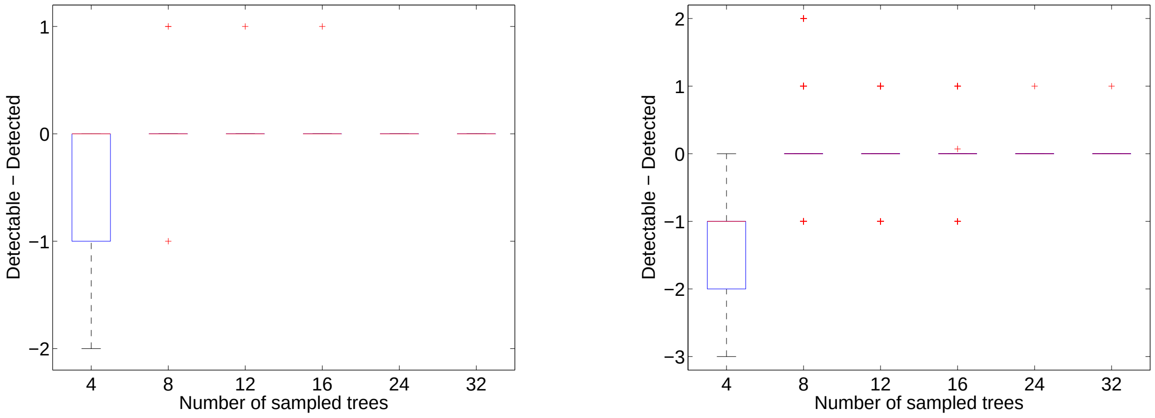
5.2. Results and Discussion

Due to space limitations, we show the results only for the 30- and 50-taxon data sets; we observed very similar trends in the cases of the 10- and 20-taxon data sets. Figure 2 shows the accuracy, in terms of the difference between detectable and detected numbers of reticulation events, of **Method M1**, while Figure 3 shows the accuracy of **Method M2**.

In the case of 30 taxa, sampled trees are selected from the network that contains ($2^5 =$) 32 different trees. There are 80 such networks, each of which was sampled with sample sizes of 2, 4, 8, 12, 16, and 24. Therefore, at each point of the x-axis in Figures 2(a) and 3(a), the result shows the distribution of the number of reticulation events for 80 different gene tree sets. Figures 2(b) and 3(b) show the results for sampled gene tree sets in 120 different networks with 50 taxa and 10 HGTs. Each of these networks contains up to 1024 different trees. From them, sampled gene tree sets are chosen with sample sizes of 2, 4, 8, 12, 16, 24, and 32. For each of the sampling sizes and each of the networks, we sampled 30 times. Therefore, at each point of the x-axis in Figures 2(b) and 3(b), the result is the distribution of the number of reticulation events for 3600 different gene tree sets.

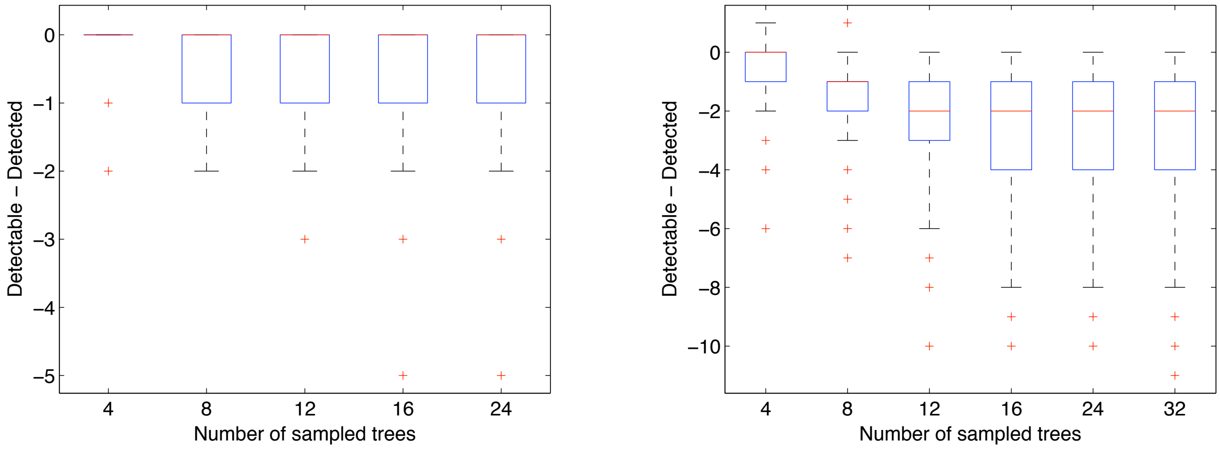
As the figures show, both methods perform very well on the 30-taxon data sets, with the median different between detectable and detected numbers of reticulations, for both methods, falling at zero. In the case of **Method M1**, there is an improvement in the accuracy as the sample size increases, as is evident from the lack of outliers and the convergence to the median value of 0. This is because, as the sample size increases, the data points become much denser so that fitting the binomial distribution becomes easier. Nonetheless, even for very sparse samples (sizes 4 and 8), the method still performs very well, as shown in Figure 2(a). **Method M2**, on the other hand, does not show clear improvement with increased sample size; to the contrary, more outliers emerge as the sample size increases (Figure 3(a)). One reason behind this is that as the sample size increases and the SPR move sets become larger, a more careful handling of the union of those sets is required than we employ in our heuristics. In some sense, this problem becomes similar to the Inclusion-Exclusion principle, where one has to avoid double-counting.

For the 50-taxon data sets, both methods also perform well, particularly **Method M1**. Even though both methods tend to overestimate the amount of reticulation in these cases (as shown by the negative values in Figures 2(b) and 3(b)), the under-estimation is very mild on average. It is worth mentioning that the results in Figure 3(b) come from much smaller sampled gene tree sets (less than 4%). From the results shown in Figures 3(a) and 3(b), sampling with size of 2, or only given a pair of gene trees, is not sufficient to estimate true number of reticulation events. For sampling with the sizes larger than 2, the results are very close to the



(a) (b)

Fig. 2. Performance of **Method M1** on the 30-taxon (a) and 50-taxon (b) data sets as a function of the sample size.



(a) (b)

Fig. 3. Performance of **Method M2** on the 30-taxon (a) and 50-taxon (b) data sets as a function of the sample size.

true number of reticulation events (5 in Figure 3(a) and 10 in Figure 3(b)) in most cases, having a difference of up to 2. **Method M2** tends to overestimate the number of reticulation events. In the worst case, the estimated results could double the actual number of reticulation events. However, the median of the distribution and the results for most cases have a converging trend when the sampling size increases.

Finally, we set out to investigate the effect of the actual sample of trees on the performance of the methods, particularly **Method M1**, since it is sensitive to the distribution of pairwise distances. For each actual network-node (reticulation event) in a simulated network N, roughly half of the trees in $\mathcal{T}(N)$ use one parent, whereas the other half use the other parent. We hypothesize that the detectability of a reticulation node is easy when half of the gene trees give signal about one of its parents, while the other half give signal about its other parent. In Figure 4, we plot the performance of **Method 1** on the 50-taxon data sets, as a function of the deviation of the trees in a sample from the balanced coverage of each reticulation event (written as "distribution deviation from 1/2" on the x-axis). Clearly, there is a correlation between the deviation from a balanced coverage of reticulation events by the trees in a sam-

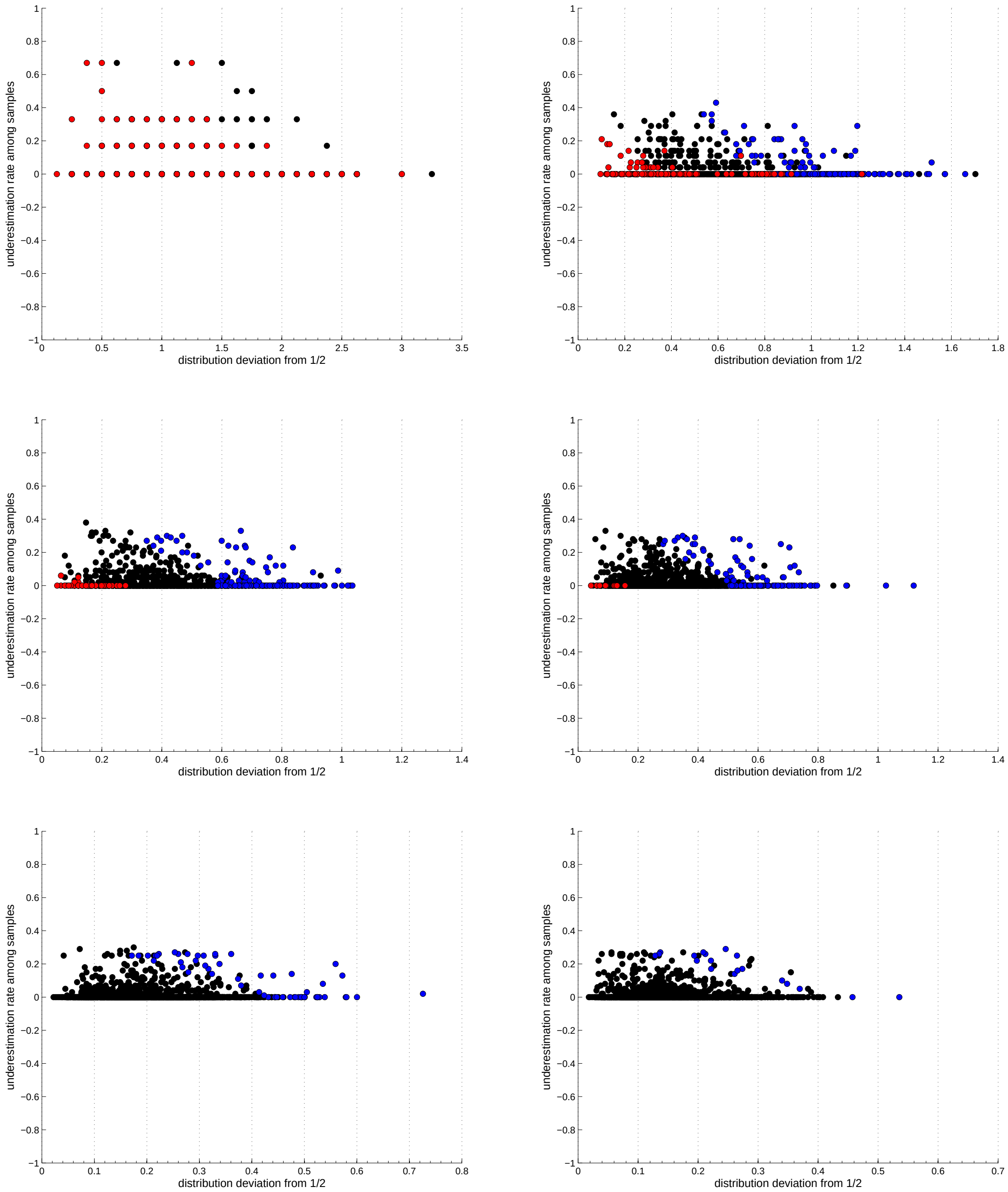


Fig. 4. Inspection of over- and under-estimation of **Method 1** as a function of the distribution deviation from 1/2 (see text for more details). Black, blue and red dots represent correct, under-, and over-estimations, respectively, of the method. Left to right, top down: sample sizes 4, 8, 12, 16, 24, and 32 (all on 50-taxon data sets).

ple and the estimation trend: over-estimations occur at lower deviation from balanced coverage, followed by correct estimation at higher deviations, and finally under-estimations occurring at the highest deviation from balanced coverage. We do not have a clear answer to why this is the case, but this leads to an interesting question about the effect of the balance of trees in a set on the detectability of reticulations.

6. CONCLUSIONS AND FUTURE RESEARCH

The increasing availability of whole-genome and multi-locus data has highlighted the need for computational tools that enable phylogenomic analyses. One such analysis entails comparing gene trees in a group of organisms, identifying their differences, and using this information to elucidate the evolutionary mechanisms that acted on the organisms during the course of their evolution. In prokaryotic organisms, it is widely believed that horizontal gene transfer (HGT) is ubiquitous, and that it plays an important role in genomic diversification.

Mathematically, the *subtree prune and regraft*, or SPR, distance between a pair of trees has been commonly used as a proxy for a lower bound on the number of HGT events, or reticulations. As a result, a wide array of mathematical results and computational tools have been developed around this distance. Nonetheless, most of these results and tools apply to a pair of trees, which is a shortcoming, particularly for phylogenomic studies involving many trees.

In this paper, we addressed the problem of estimating the amount of reticulation that is detectable in a collection of gene trees, assuming all incongruence among the trees is due to reticulate evolution (i.e., ruling out any other discord processes, such as incomplete lineage sorting, gene duplication/loss, etc.). We provided two algorithmic strategies for this task, both of which showed promising results in simulations.

Our main task for future research is to apply these strategies to real data, not only to assess the performance of the methods, but also to better understand reticulate evolution in prokaryotes. However, in addition to going beyond two trees, a major challenge needs to be addressed in order to apply tools to real data. While almost all results and tools developed for the SPR distance problem assume the trees have the same leaf labels, this may not be the case in phylogenomic studies. In particular, incomplete taxon sampling and disparity in sequence coverage for different organisms may result in "missing" genes for some organisms. Biologically, gene duplication and loss may result in multiple or no copies of certain genes in some organisms. Further, a horizontal gene transfer event from outside the group of organisms under study may give rise to genes that are present in some, but not all, of the organisms. Last but not least, HGT events across genes may not be independent, as a single event may result in the transfer of a large genomic region that contains multiple genes. All these issues need to be addressed in order to facilitate a true phylogenomic study; otherwise, analyses would have to be restricted to a small fraction of the genomic data, rendering their results and conclusions unreflective of the true, global picture [28].

7. ACKNOWLEDGMENTS

This work was supported in part by the Rice Terascale Cluster funded by NSF under grant EIA-0216467, Intel, and HP, as well as grants CCF-0622037 (NSF) and R01LM009494 from the National Library of Medicine to Luay Nakhleh. The contents are solely the responsibility of the authors and do not necessarily represent the official views of the DOE, NSF, National Library of Medicine or the National Institutes of Health.

References

1. B. Gemeinholzer, "Phylogenetic networks," in *Analysis of Biological Networks* (B. H. Junker and F. Schreiber, eds.), pp. 255–282, John Wiley and Sons Ltd, 2008.
2. D. Huson, "Split networks and reticulate networks," in *Reconstructing Evolution, New Mathematical and Computational Advances* (O. Gascuel and M. Steel, eds.), pp. 247–276, Oxford University Press, 2007.
3. D. Huson and D. Bryant, "Application of phylogenetic networks in evolutionary studies," *Molecular Biology and Evolution*, vol. 23, no. 2, pp. 254–267, 2006.
4. C. Linder, B. Moret, L. Nakhleh, and T. Warnow, "Network (reticulate) evolution: Biology, models, and algorithms," in *The Pacific Symposium on Biocomputing*, 2004.
5. V. Makarenkov, D. Kevorkov, and P. Legendre, "Phylogenetic network construction approaches," in *Applied Mycology and Biotechnology*, pp. 61–97, 2006.
6. D. Morrison, "Networks in phylogenetic analysis: new tools for population biology," *International Journal of Parasitology*, vol. 35, pp. 567–582, 2005.
7. L. Nakhleh, "Evolutionary phylogenetic networks: Models and issues," in *The Problem Solving Handbook for Computational Biology and Bioinformatics*

(L. Heath and N. Ramakrishnan, eds.), Springer, 2010. To appear.

8. W. Maddison, "Gene trees in species trees," *Systematic Biology*, vol. 46, no. 3, pp. 523–536, 1997.
9. B. Allen and M. Steel, "Subtree transfer operations and their induced metrics on evolutionary trees," *Annals of Combinatorics*, vol. 5, pp. 1–13, 2001.
10. M. Baroni, S. Grunewald, V. Moulton, and C. Semple, "Bounding the number of hybridisation events for a consistent evolutionary history," *J. Math. Biol.*, vol. 51, pp. 171–182, 2005.
11. P. Humphries and C. Semple, "Note on the hybridization number and subtree distance in phylogenetics," *Applied Mathematics Letters*, vol. 22, no. 4, pp. 611–615, 2009.
12. M. Bordewich and C. Semple, "On the computational complexity of the rooted subtree prune and regraft distance," *Annals of Combinatorics*, vol. 8, pp. 409–423, 2004.
13. Y. Wu, "A practical method for exact computation of subtree prune and regraft distance," *Bioinformatics*, vol. 25, no. 2, pp. 190–196, 2009.
14. M. Hallett and J. Lagergren, "Efficient algorithms for lateral gene transfer problems," in *Proc. 5th Ann. Int'l Conf. Comput. Mol. Biol. (RECOMB01)*, pp. 149–156, 2001.
15. L. Nakhleh, D. Ruths, and L. Wang, "RIATA-HGT: A fast and accurate heuristic for reconstrucing horizontal gene transfer," in *Proceedings of the Eleventh International Computing and Combinatorics Conference (COCOON 05)* (L. Wang, ed.), pp. 84–93, 2005. LNCS #3595.
16. C. Than, D. Ruths, and L. Nakhleh, "PhyloNet: A software package for analyzing and reconstructing reticulate evolutionary relationships," *BMC Bioinformatics*, vol. 9, p. 322, 2008.
17. R. Beiko and N. Hamilton, "Phylogenetic identification of lateral genetic transfer events," *BMC Evolutionary Biology*, vol. 6, p. 15, 2006.
18. D. MacLeod, R. Charlebois, F. Doolittle, and E. Bapteste, "Deduction of probable events of lateral gene transfer through comparison of phylogenetic trees by recursive consolidation and rearrangement," *BMC Evolutionary Biology*, vol. 5, p. 27, 2005.
19. P. Goloboff, "Calculating SPR distances between trees," *Cladistics*, vol. 24, pp. 591–597, 2007.
20. D. Huson and R. Rupp, "Summarizing multiple gene trees using cluster networks," in *Proceedings of the Workshop on Algorithms in Bioinformatics* (K. Crandall and J. Lagergren, eds.), vol. 5251 of *Lecture Notes in Bioinformatics*, pp. 296–305, 2008.
21. R. Beiko and M. Ragan, "Untangling hybrid phylogenetic signals: Horizontal gene transfer and artifacts of phylogenetic reconstruction," *Methods Mol Biol.*, vol. 532, pp. 241–256, 2009.
22. S. Kullback, "The Kullback-Leibler distance," *The American Statistician*, vol. 41, pp. 340–341, 1987.
23. M. Suchard, "Stochastic models for horizontal gene transfer: taking a random walk through tree space," *Genetics*, vol. 170, pp. 419–431, 2005.
24. C. Than, D. Ruths, H. Innan, and L. Nakhleh, "Confounding factors in HGT detection: Statistical error, coalescent effects, and multiple solutions," *Journal of Computational Biology*, vol. 14, no. 4, pp. 517–535, 2007.
25. C. Than and L. Nakhleh, "SPR-based tree reconciliation: Non-binary trees and multiple solutions," in *Proceedings of the Sixth Asia Pacific Bioinformatics Conference (APBC)*, pp. 251–260, 2008.
26. A. Rambaut, "Phylogen: Phylogenetic tree simulator package," 2002. Available from http://tree.bio.ed.ac.uk/software/phylogen/.
27. N. Galtier, "A model of horizontal gene transfer and the bacterial phylogeny problem," *Systematic Biology*, vol. 56, no. 4, pp. 633–642, 2007.
28. T. Dagan and W. Martin, "The tree of one percent," *Genome Biology*, vol. 7, no. 10, p. 118, 2006.