JOURNAL OF COMPUTATIONAL BIOLOGY
Volume 14, Number 4, 2007
© Mary Ann Liebert, Inc.
Pp. 517–535
DOI: 10.1089/cmb.2007.A010

Confounding Factors in HGT Detection: Statistical Error, Coalescent Effects, and Multiple Solutions

CUONG THAN,[1] DEREK RUTHS,[1] HIDEKI INNAN,[2] and LUAY NAKHLEH[1]

ABSTRACT

Prokaryotic organisms share genetic material across species boundaries by means of a process known as *horizontal gene transfer* (HGT). This process has great significance for understanding prokaryotic genome diversification and unraveling their complexities. Phylogeny-based detection of HGT is one of the most commonly used methods for this task, and is based on the fundamental fact that HGT may cause gene trees to disagree with one another, as well as with the species phylogeny. Using these methods, we can compare gene and species trees, and infer a set of HGT events to reconcile the differences among these trees. In this paper, we address three factors that confound the detection of the true HGT events, including the donors and recipients of horizontally transferred genes. First, we study experimentally the effects of error in the estimated gene trees (statistical error) on the accuracy of inferred HGT events. Our results indicate that statistical error leads to overestimation of the number of HGT events, and that HGT detection methods should be designed with unresolved gene trees in mind. Second, we demonstrate, both theoretically and empirically, that based on topological comparison alone, the number of HGT scenarios that reconcile a pair of species/gene trees may be exponential. This number may be reduced when branch lengths in both trees are estimated correctly. This set of results implies that in the absence of additional biological information, and/or a biological model of how HGT occurs, multiple HGT scenarios must be sought, and efficient strategies for how to enumerate such solutions must be developed. Third, we address the issue of lineage sorting, how it confounds HGT detection, and how to incorporate it with HGT into a single stochastic framework that distinguishes between the two events by extending population genetics theories. This result is very important, particularly when analyzing closely related organisms, where coalescent effects may not be ignored when reconciling gene trees. In addition to these three confounding factors, we consider the problem of enumerating all valid coalescent scenarios that constitute plausible species/gene tree reconciliations, and develop a polynomial-time dynamic programming algorithm for solving it. This result bears great significance on reducing the search space for heuristics that seek reconciliation scenarios. Finally, we show, empirically, that the locality of incongruence between a pair of trees has an impact on the numbers of HGT and coalescent reconciliation scenarios.

Key words: phylogeny, horizontal gene transfer, coalescent.

[1]Department of Computer Science, Rice University, Houston, Texas.
[2]School of Advanced Sciences, Graduate University for Advanced Studies, Kanagawa, Japan.

1. INTRODUCTION

WHEREAS EUKARYOTES EVOLVE mainly though lineal descent and mutations, bacteria obtain a large proportion of their genetic diversity through the acquisition of sequences from distantly related organisms via horizontal gene transfer (HGT) (Doolittle et al., 2003; Ochman et al., 2000). There has been a large "ideological and rhetorical" gap between researchers who believe that HGT is so rampant that a prokaryotic phylogenetic tree is useless and those who believe that HGT is mere "background noise" which does not affect the reconstructibility of a phylogenetic tree for bacterial genomes. Supporting arguments for these two views have been published. For example, the heterogeneity of genome composition between closely related strains (e.g., in *Escherichia coli* only 40% of genes are shared in common by three *E. coli* strains [Welch et al., 2002]) supports the former view, whereas the well-supported phylogeny reconstructed by Lerat et al. (2003) from about 100 "core" genes in γ-Proteobacteria gives evidence in favor of the latter view. Nonetheless, regardless of the views and the accuracy of the various analyses, there is a consensus as to the occurrence of HGT and the evolutionary role it plays in bacterial genome diversification. Further, HGT is a main process by which bacteria develop resistance to antibiotics (Paulsen et al., 2003), is considered a primary explanation of incongruence among gene phylogenies, and is a significant obstacle to reconstructing the Tree of Life (Daubin et al., 2003).

The HGT detection problem concerns the detection of the genes that are horizontally transferred into the genome, the donors and recipients of every horizontally transferred gene, and the number of HGT events that occurred during the evolutionary history of a set of species. When HGT occurs, the evolutionary history of the gene(s) involved does not necessarily agree with that of the species phylogeny. This observation is the fundamental basis of the phylogeny-based HGT detection approach: trees for individual genes are reconstructed (and sometimes a species tree is reconstructed as well, using other data), and their disagreements are identified to estimate the number (how many) as well as locations (donors and recipients) of HGT events. Beside the computationally challenging problem of quantifying disagreements among trees for the sake of detecting HGT, major challenges that face this approach include (1) determining whether the disagreements are indeed due to HGT, and (2) whether there is a unique HGT "scenario." Yet, these two challenges encompass a host of issues of which we address three. First, since trees are at best partially known, they have to be reconstructed using a phylogeny reconstruction method. We investigate the impact that the quality of reconstructed trees has on HGT detection. Second, under the assumption that HGT is actually the source of tree disagreements, we investigate the uniqueness of a solution to the HGT detection problem, and establish bounds on the number of possible minimal HGT scenarios. Finally, among closely related species, *lineage sorting* due to random genetic drift may also cause tree incongruence, thus mimicking the effects of HGT on phylogenies. In this case, accurate HGT detection requires determining the actual cause of tree incongruities, and making the appropriate reconciliation. We make preliminary progress on incorporating HGT into the coalescent model, so as to produce a stochastic framework for classifying population-level events (such as lineage sorting) and species-level events (such as HGT).

In addition to these three confounding factors, we consider the problem of enumerating all valid coalescent scenarios that constitute plausible species/gene tree reconciliations, and develop a polynomial dynamic programming algorithm for solving it. This result bears a great significance on reducing the solution space for heuristics that search for reconciliation scenarios. Finally, we show, empirically, that the locality of incongruence between a pair of trees has an impact on the numbers of HGT and coalescent reconciliation scenarios.

We draw several conclusions from this work. First, to obtain accurate estimates of HGT-based tree incongruence, poorly supported edges of reconstructed trees should be removed. Though an important task to conduct, removing (or contracting) poorly supported edges is not a straightforward task, since standard methods that are in common use for estimating branch support, such as bootstrapping and posterior probabilities in Bayesian analyses, have been shown to be overly "conservative" or "liberal" under various circumstances (Ruths and Nakhleh, 2006). Second, eliminating statistical error from reconstructed trees leads to non-binary trees, and hence phylogeny-based HGT detection methods should be designed to handle such trees (rather than focus on binary trees, which many existing tools do). Third, more than one maximally parsimonious solution (a solution that has the minimum number of HGT edges, or events, to explain the species and gene tree incongruence) may exist, and hence HGT detection methods should search for all such

solutions, unless additional biological information is given, or a model that simultaneously incorporates HGT, speciation, and extinction events (Kunin and Ouzounis, 2003). Finally, trees may be incongruent due to processes other than HGT; hence, classifying the sources of incongruence and reconciling them accordingly is imperative.

2. TREE INCONGRUENCE AND HGT DETECTION

A gene tree is a model of how a gene evolves. As a gene at a locus in the genome replicates and its copies are passed on to more than one offspring, branching points are generated in the gene tree. Because the gene has a single ancestral copy, barring recombination, the resulting history is a branching tree (Maddison, 1997). Thus, within a species, many tangled gene trees can be found, one for each nonrecombined locus in the genome. Exploring incongruence among gene trees is the basis for phylogeny-based HGT detection and reconstruction.

We illustrate some of the scenarios that may lead to gene tree incongruence in Figure 1. The species tree is represented by the "tubes"; it has A and B as sister taxa whose most recent common ancestor (MRCA) is a sister taxon of C.

In the case of HGT, shown in Figure 1a, genetic material is transferred from one lineage to another. Sites that are not involved in a horizontal transfer are inherited from the parent while other sites are horizontally transferred from another species. Figure 1b gives an example of a gene tree that disagrees with the species phylogeny because of lineage sorting due to random genetic drift: the genes of B and C coalesced before their MRCA coalesced with the gene of species A. Moreover, sometimes multiple events "cancel out" one another's effects when co-occurring in the same dataset; for example, in Figure 1c, lineage sorting "hides" the incongruence between the species and gene trees (tree topologies) that would have resulted from the HGT event. Another factor that may lead to gene and species tree disagreements is that trees reconstructed by phylogenetic methods may not be completely accurate (we refer to this as *statistical error* in the trees); hence, disagreements among trees due to such inaccuracies may trigger HGT "signal," thus leading to overestimation of the actual HGT events.

Notice that in the case of lineage sorting, the species phylogeny is still a tree, and the gene trees should be reconciled within its branches. However, in the case of HGT, the evolutionary history of the species genomes may not be represented by phylogenetic trees; rather, *phylogenetic networks* are the appropriate model (Moret et al., 2004; Kunin et al., 2005). The phylogeny-based HGT detection problem seeks the phylogenetic network with minimum number of *reticulation nodes*, e.g., HGT edges, to reconcile the species and gene trees. The minimization simply reflects a maximally parsimonious solution: in the absence of any additional biological knowledge, the simplest solution is sought. In the case, the simplest solution is one that invokes the minimum number of HGT events to explain tree incongruence. There has been a large body of work on this problem (Hallett and Lagergren, 2001; Nakhleh et al., 2004; Bordewich and Semple, 2005; Nakhleh et al., 2005; Makarenkov, 2001).

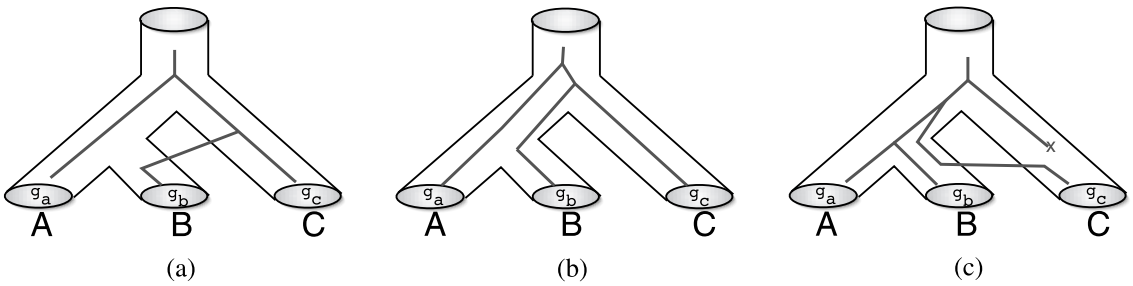

FIG. 1. Gene tree that disagrees with the species tree due to HGT from C to B **(a)** and lineage sorting due to random genetic drift **(b)**. **(c)** The effect of the HGT event (from B to C) is "canceled out" by random genetic drift, resulting in congruent species and gene trees.

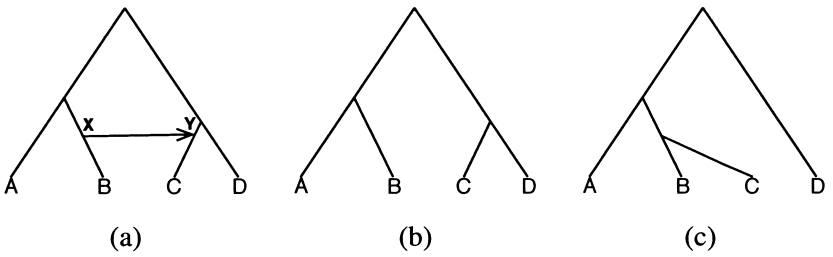

FIG. 2. **(a)** A phylogenetic network with a single HGT event from X to Y. **(b)** The underlying organismal (species) tree. **(c)** The tree of a horizontally transferred gene.

2.1. Terminology and definitions

Let $T = (V, E)$ be a tree, where V and E are the *tree nodes* and *tree edges*, respectively, and let $\mathcal{L}(T)$ denote its leaf set. Further, let $\mathcal{X}$ be a set of taxa (species). Then, T is a phylogenetic tree over $\mathcal{X}$ if there is a bijection between $\mathcal{X}$ and $\mathcal{L}(T)$. A tree T is said to be *rooted* if the set of edges E is directed and there is a single distinguished internal vertex r with in-degree 0. A phylogenetic network $N = N(T) = (V', E')$ over the taxa set $\mathcal{X}$ is derived from $T = (V, E)$ by adding a set Ξ of edges to T, where each edge $h \in \Xi$ is added as follows: (1) split an edge $e \in E$ by adding new node, v_e; (2) split an edge $e' \in E$ by adding new node, $v_{e'}$; (3) finally, add a directed *HGT edge* from v_e to $v_{e'}$. In this case, we write $N = T + \Xi$. Figure 2a shows a phylogenetic network obtained by adding a single HGT edge to the tree in Figure 2b.

In a case where horizontal transfer of a single gene is involved, the edges e' that are split in step (2) above must be unique. In other words, no more than a single HGT edge may be incident into a single tree edge. However, this is not necessarily true if the phylogenetic network models the evolutionary history of multiple genes. In this case, an edge e' may be, for example, split twice, once because of an HGT involving gene g and another because of an HGT involving gene g'.

It is important to note that our definition of a phylogenetic network allows adding an HGT edge from a tree edge e to another tree edge e' "below" it. While this seems to violate biological constraints (such as the temporal co-existence of the donor and recipient), this case may arise in practice due, for instance, to incomplete taxon sampling or extinction. For a more thorough discussion of this issue, and modeling reticulate evolution in general, the reader is referred to Moret et al. (2004).

Finally, we denote by $\mathcal{T}(N)$ the set of all trees contained inside network N. Each such tree is obtained by the following two steps: (1) for each node of in-degree 2, remove one of the incoming edges, and then (2) for every node x of in-degree and out-degree 1, whose parent is u and child is v, remove node x and its two adjacent edges, and add a new edge from u to v. For example, the two trees in Figure 2b,c are the only members of $\mathcal{T}(N)$, where N is the network in Figure 2a.

Definition 1. (The HGT Reconstruction Problem)

Input: *Species tree ST and gene tree GT.*
Output: *Minimum-cardinality set* Ξ *of HGT edges, such that* $N = ST + \Xi$ *and* $GT \in \mathcal{T}(N)$.

3. THE EFFECT OF STATISTICAL ERROR ON HGT DETECTION

In this section we investigate, through simulations, the effect of error in the reconstructed trees on the detection of HGT. In particular, we consider the minimum number of HGT events inferred by HGT detection methods, as well as the number of such maximally parsimonious solutions found by these methods.

Experimental settings. We used the `r8s` tool (Sanderson, n.d.) to generate four random birth-death phylogenetic trees, T_i, $i \in \{10, 25, 50, 100\}$, where i denotes the number of taxa in the tree. The `r8s` tool generates molecular clock trees; we deviated the trees from this hypothesis by multiplying each edge in the tree by a number randomly drawn from an exponential distribution. The expected evolutionary diameter

(longest path between any two leaves in the tree) is 0.2. Then, from each model "species" tree T_i, we generated five different "gene" trees, $T_{i,j}$, $j \in \{1, 2, 3, 4, 5\}$, where j denotes the number of *subtree prune and regraft* (SPR) moves applied to T_i to obtain $T_{i,j}$.[1] The SPR moves were applied to simulated HGT events such that no cycles are created, and such that there is no redundancy (i.e., if k SPR moves were simulated to obtain T_2 from T_1, then T_2 cannot be obtained from T_1 be fewer than k SPR moves). Beside these two requirements, the choice of the "donor" and "recipient" (i.e., the source and target of an HGT edge) were done purely randomly, without considerations to evolutionary distance between the two endpoints, or any other issues.

For each T_i and $T_{i,j}$, $i \in \{10, 25, 50, 100\}$ and $j \in \{1, 2, 3, 4, 5\}$, and for each sequence length $\ell \in \{250, 500, 1000, 2000, 4000, 8000\}$, we generated 30 DNA sequence alignments $S_i^{\ell}[k]$ and $S_{i,j}^{\ell}[k]$, $1 \leq k \leq 30$, whose evolution was simulated down their corresponding trees under the GTR+Γ+I (gamma distributed rates, with invariable sites) model of evolution, using the Seq-gen tool (Rambaut and Grassly, 1997). We used the parameter settings of Zwickl and Hillis (2002). Then, from each sequence alignment, we reconstructed a tree *TNJ* using the Neighbor Joining (NJ) method (Saitou and Nei, 1987), and another tree using a maximum parsimony heuristic as implemented in PAUP* (Swofford, 1996). Since the maximum parsimony heuristic may return a set of optimal trees, for each alignment we only considered the *strict consensus* of each such set, and referred to that as the tree *TMP*. At the end of this process we had 4 trees T_i, 20 trees $T_{i,j}$, 720 NJ trees $TNJ_i^{\ell}[k]$, 3600 NJ trees $TNJ_{i,j}^{\ell}[k]$, 720 MP trees $TMP_i^{\ell}[k]$, and 3600 MP trees $TMP_{i,j}^{\ell}[k]$ ($i \in \{10, 25, 50, 100\}$, $j \in \{1, 2, 3, 4, 5\}$, $1 \leq k \leq 30$, and $\ell \in \{250, 500, 1000, 2000, 4000, 8000\}$).

To compute minimal HGT scenarios as well as the number of such scenarios, we applied two methods to pairs of species and gene trees: LatTrans (Hallett and Lagergren, 2001; Addario-Berry et al., 2003) and RIATA-HGT (Nakhleh et al., 2005), which has been recently extended and improved to compute multiple minimal solutions (Jin et al., 2007; the original method is described in Nakhleh et al., 2005), computes only a single minimal solutions, since the emphasis of the underlying algorithm was on estimating the minimum number of HGT events). There are several methods for recovering candidate HGT edges based on comparing a pair of trees (Hallett and Lagergren, 2001; Makarenkov, 2001; Addario-Berry et al., 2003; Nakhleh et al., 2005; MacLeod et al., 2005; Beiko and Hamilton, 2006). In this work, we did not intend to study or compare the performance of these methods, but rather to try to quantify and understand the effect of various factors on the estimation of the number of HGT events. For this purpose, we chose two different methods: LatTrans and RIATA-HGT. Since both methods are heuristics, independently developed, and their relative performance is unknown (in terms of accuracy), we used both to ensure that the effects measured reflect general trends, rather than issues specific to a particular heuristic. Indeed, the similar trends observed in the experiments raise certain points that seem to be independent of the heuristic used.

Both tools were applied to three different types of pairs of trees.

Type I pairs $(T_i, T_{i,j})$: in this case, the species and gene trees are assumed to be correct.

Type II pairs $(T_i, TNJ_{i,j}^{\ell}[k])$ and $(T_i, TMP_{i,j}^{\ell}[k])$: in this case, the species tree is correct, and the gene trees are estimated (using NJ and MP, respectively).

Type III pairs $(TNJ_i^{\ell}[k], TNJ_{i,j}^{\ell}[k])$ and $(TMP_i^{\ell}[k], TMP_{i,j}^{\ell}[k])$: in this case, both the species and gene trees are inferred.

The goal of running the methods in these different ways is to estimate the error due to inaccuracy in the different trees. Due to space limitations, we only show results using NJ trees, 25-taxon trees (Since LatTrans cannot handle non-binary trees, it was not run on MP trees, and from RIATA-HGT's results on MP trees, the trends on MP trees are very similar). In each run of a tool on a pair of trees, we computed two values: the number of inferred HGT events, and the number of such scenarios (or solutions) found by the method. In Type II and Type III pairs, we report the average of all 30 runs for each combination of i, j, and ℓ.

[1] An SPR move simulates an HGT event.

3.1. The effect of statistical error on estimating the number of HGT events

Both LatTrans and RIATA-HGT computed the correct number of SPR moves (i.e., HGT edges) when applied to Type I pairs. In other words, when both the species and gene trees were correct, both methods made an accurate estimation of the number of HGT events. The performance of both methods, in terms of the number of inferred HGT events, on Type II and Type III pairs of trees is shown in Figure 3. Figures 3a and 3b show that, when the species tree is accurate, and the gene tree is inferred, both methods accurately estimate the number of HGT events for the case of five HGT events when the sequences are of length 8000. They overestimate the number for all other cases, at all sequence lengths. As the sequence length increases, the trees inferred by NJ become more accurate, since NJ is *statistically consistent* (Atteson, 1999), and hence the improvement in the performance of the methods as the sequence length increases. At sequence length 250, the methods have the worst performance. When both the species and gene trees are inferred, the overestimation becomes larger, as shown in Figures 3c and 3d. In this case, even at sequence length 8000 the methods do overestimate the actual number of HGT events. It is worth noting that both methods have almost identical performance in terms of the number of HGT events inferred (RIATA-HGT does slightly better in some cases at sequence length 1000). However, RIATA-HGT is orders of magnitude faster. Figure 4 shows the relative performance (in terms of actual running time) of the two methods on 25-taxon NJ trees. Specially, LatTrans took several days on each pair of 50-taxon trees, and for sequence length 250 it crashed after 4 days without returning results.

Given that the two methods accurately estimated the number of HGT events in Type I pairs of trees, i.e., accurate species and gene trees, the results show that error in inferred trees (one or both) leads to overestimation of the number of HGT events. The overestimation is even larger for the larger data

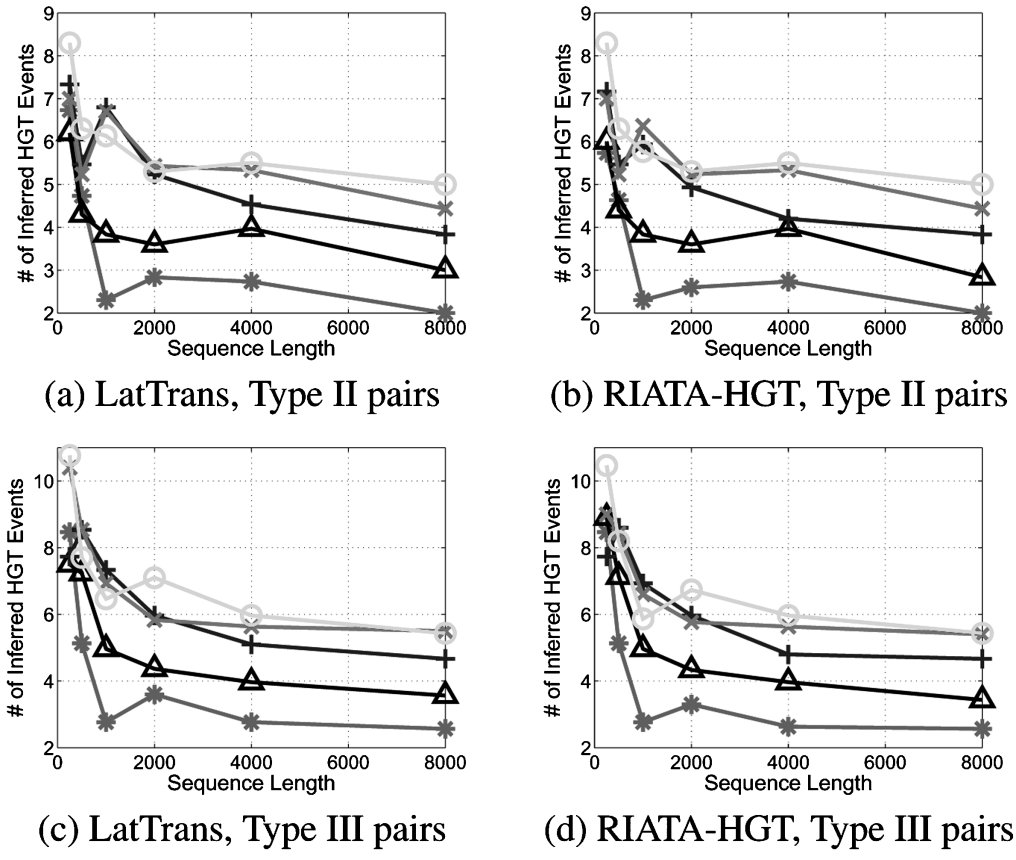

FIG. 3. The number of HGT events inferred by LatTrans and RIATA-HGT, as a function of the sequence length. Each curve corresponds to one of the five actual numbers of HGT events: $\star$, one HGT; $\triangle$, two HGTs; $+$, three HGTs; $\times$, four HGTs; and $\circ$, five HGTs. 25-taxon trees inferred using NJ.

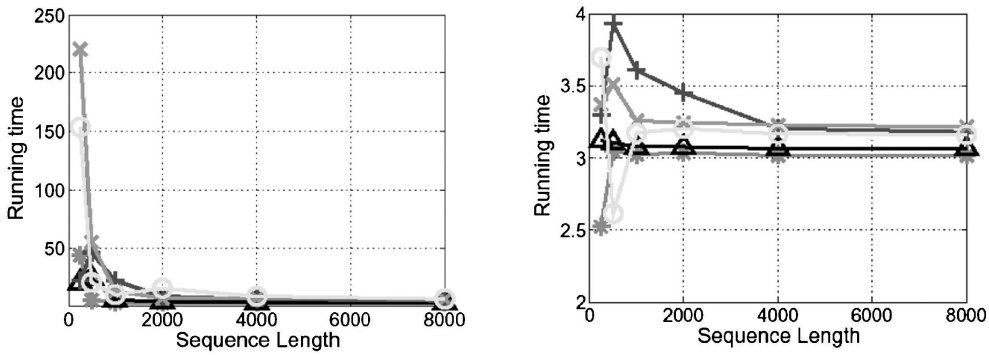

FIG. 4. The performance of LatTrans and RIATA-HGT in terms of actual running times (in seconds), as functions of the sequence length. Each curve corresponds to one of the five actual numbers of HGT events: ⋆, one HGT; △, two HGTs; +, three HGTs; ×, four HGTs; and ∘, five HGTs.

sets (50- and 100-taxon trees). Therefore, it is important to eliminate statistical error from trees before estimating HGT events. Ruths and Nakhleh (2006) have studied the performance of various methods for eliminating incorrect edges while maintaining accurate ones. This elimination, in the form of contracting poorly supported edges, may lead to non-binary trees in certain cases, which cannot be handled by LatTrans, although they can be handled by RIATA-HGT.

4. THE UNIQUENESS OF HGT SCENARIOS

Moret et al. (2004) showed that a phylogenetic network that reconciles two trees need not be unique, by showing two phylogenetic networks, each with a single reticulation event, that reconcile the same pair of trees. Further, they showed how branch lengths could be used to resolve the non-uniqueness question in this simple case. Here we show that the number of possible maximally parsimonious (with minimum number of HGT events) phylogenetic networks that reconcile a pair of trees may actually be exponential. Further, we discuss when branch lengths may not be sufficient to resolve the non-uniqueness issue.

The number of maximally parsimonious HGT scenarios that reconcile a pair of trees (species and gene trees, for example) may be exponentially large, as illustrated in Figure 5. The species and gene trees in the figure, ST and GT, respectively, contain $3k$ leaves and differ in that X_{i2} is closer to X_{i1} than to X_{i3} in tree ST, and closer to X_{i3} than to X_{i1} in tree GT, for $1 \leq i \leq k$. For every triplet $\langle X_{i1}, X_{i2}, X_{i3} \rangle$ of taxa, one of three HGT edges is needed to reconcile the difference in topologies of the triplet based on the two trees ST and GT: (1) the edge $H_{i1} : X_{i3} \rightarrow X_{i2}$, (2) the edge $H_{i2} : X_{i2} \rightarrow X_{i3}$, or (3) the edge $H_{i3} : m_i \rightarrow X_{i1}$, where m_i is the edge incoming into the most recent common ancestor (node) of the triplet of taxa; these three scenarios are shown in Figure 6. To reconcile the differences among all k

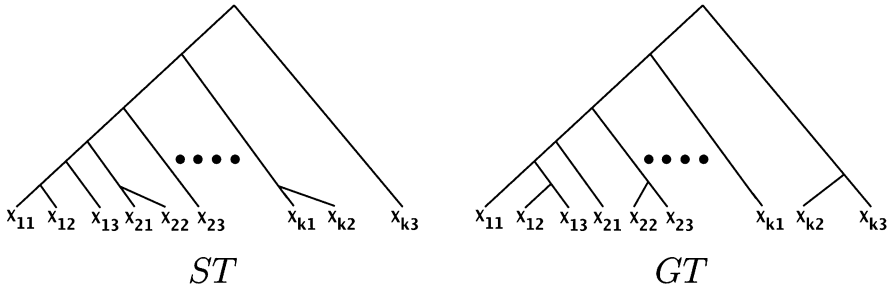

FIG. 5. A species tree ST and a gene tree GT with $3k$ leaves. The two trees differ in k places: the species tree has X_{i1} and X_{i2} as siblings, whereas the gene tree has X_{i2} and X_{i3} as siblings ($1 \leq i \leq k$). There are 3^k maximally parsimonious HGT scenarios that reconcile the two trees.

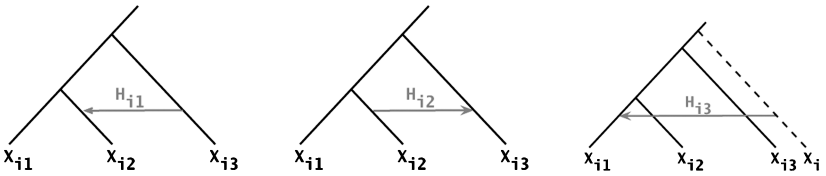

FIG. 6. The three possible scenarios for reconciling the topologies of the triplet $\langle X_{i1}, X_{i2}, X_{i3}\rangle$ based on the species and gene trees, *ST* and *GT*, respectively, in Figure 5.

triplets, there are 3^k HGT scenarios, since there are k triplets to reconcile, and for each triplet there are three possible reconciliations. Two observations are in order. First, since the donor and recipient of a gene have to co-exist in time (Moret et al., 2004), and given that the topology of a phylogeny defines a partial order on the set of extant and ancestral taxa (ancestral taxa *precede* their descendants in this partial order), it follows that edge H_{i3} can be part of an HGT solution only if certain taxa went extinct or were not sampled. This case is illustrated in Figure 6, where the dashed line represents the lineage for taxon X_i which is not present in the set of taxa under consideration but whose existence must be invoked to explain the HGT edge H_{i3}.

Let δ_{ST} and δ_{GT} be the pairwise distance matrices of the set of taxa based on the species and gene trees *ST* and *GT*, respectively, in Figure 5, and let us consider the triplet of taxa in Figure 6. There are three cases. (1) The scenario H_{i1} is plausible if and only if $\delta_{ST}(X_{i1}, X_{i3}) \approx \delta_{GT}(X_{i1}, X_{i3})$ and $\delta_{ST}(X_{i1}, X_{i2}) \not\approx \delta_{GT}(X_{i1}, X_{i2})$. (2) The scenario H_{i2} is plausible if and only if $\delta_{ST}(X_{i1}, X_{i2}) \approx \delta_{GT}(X_{i1}, X_{i2})$. (3) The scenario H_{i3} is plausible if and only if If $\delta_{ST}(X_{i2}, X_{i3}) \approx \delta_{GT}(X_{i2}, X_{i3})$. Since the conditions in the three cases are mutually exclusive, it follows the branch lengths, when estimated accurately, can be used to correctly resolve the non-uniqueness issue in this case. However, estimating branch lengths to a high degree of accuracy such that the above three cases are distinguished accurately is a very challenging task. Further, even if branch lengths are estimated accurately, if the evolutionary distance between the donor and recipient is very small, distinguishing among the cases becomes more challenging.

4.1. On the number of minimal HGT scenarios: theoretical results

In this section, we derive an upper bound on the number of minimal HGT scenarios for reconciling two trees.

Given two trees *ST* and *GT*, we denote by $\kappa^{ST,GT}$ the number of HGT edges in any solution to the HGT Reconstruction Problem, and by $\mathcal{N}^{ST,GT}$ the set of all solutions. When the context is clear, we omit the tree names from the superscript.

Given a species tree *ST* and a gene tree *GT*, each with n leaves, there are $O(n^2)$ different HGT edges that can be added to it. If the cardinality of a minimal solution to the HGT Reconstruction Problem on the pair (ST, GT) is k, then there can be at most $O(n^{2k})$ solutions of size k. We now provide a tighter upper bound on the number of solutions, and show a pair of trees for which this upper bound is exact.

Theorem 1. *For any pair of trees, ST and GT, we have*

$$|\mathcal{N}| \leq 3^{\kappa}.$$

Proof. We prove the theorem by induction on κ. For the base case, let $\kappa = 1$. Then, there are two subtrees t_1 and t_2 that are siblings in *GT* but not siblings in *ST*. There are two cases: (1) on the undirected path from the root of t_1 to t_2 there are exactly two nodes (excluding the roots of t_1 and t_2), or (2) on that path, excluding the roots, there are more than two nodes. Notice that there has to be at least one node on the path, which is the least common ancestor of the two subtrees. In the first case, there is exactly one subtree t' that lies between t_1 and t_2 in the species tree (Fig. 7). There are three possible ways, in this case, to make t_1 and t_2 siblings, and these correspond to the three HGT edges (numbered 1, 2, and 3) in Figure 7. Hence, there are three solutions to the HGT reconstruction problem. In the second case, only one solution is possible, since the location of the node in *GT* whose two children are the roots of t_1 and

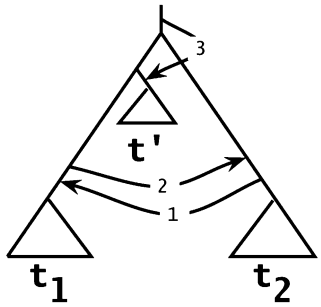

FIG. 7. Illustration of case (1) in the base case of the inductive proof of Theorem 1. The tree is denoted by the solid lines, and its leaves are the leaves of subtrees t_1, t_2, and t_3. Three phylogenetic networks, each of which contains a single HGT edge and induces the same pair of trees, are obtained by adding one of the tree HGT edges, denoted by arrows 1, 2, and 3. While HGT edge 3 may indicate violation of the temporal co-existence of the donor and recipient of a gene, this may be possible in practice due to, for instance, incomplete taxon sampling or extinction (for further details, see Section 2.1).

t_2 is fixed, given that there is more than one subtree between these two subtrees. Hence, for $\kappa = 1$, we have $|\mathcal{N}| \leq 3$.

For the induction step, assume the theorem holds for any pair of trees T_1 and T_2, where $\kappa^{T_1,T_2} < k$, and let ST and GT be two trees such that $\kappa^{ST,GT} = k$. Then, there exists a tree T' such that $\kappa^{ST,T'} = k-1$ and $\kappa^{T',GT} = 1$. By the induction hypothesis, we have $|\mathcal{N}^{ST,T'}| \leq 3^{k-1}$ and $|\mathcal{N}^{T',GT}| \leq 3$. Any network $N \in \mathcal{N}(ST, GT)$ can be written as $N = ST + (\Xi' + \{h\})$, where $ST + \Xi' \in \mathcal{N}^{ST,T'}$ and $ST + \{h\} \in \mathcal{N}^{T',GT}$. Hence, $|\mathcal{N}^{ST,GT}| \leq 3^{k-1} \cdot 3^1 = 3^k$. ■

Figure 5 shows two trees ST and GT where $|\mathcal{N}| = 3^\kappa$; i.e., the two trees achieve the maximum number of solutions.

4.2. On the number of minimal HGT scenarios: empirical results

In our simulation study (using the same experimental setup described in Section 3), we looked at the number of maximally parsimonious solutions that were computed by LatTrans and RIATA-HGT; the results for 25-taxon NJ trees are shown in Figure 8. All fours graphs show that, regardless of whether the actual or inferred species trees are used, both methods estimate a large number of maximally parsimonious solutions. The figures show that the number decreases as the sequences used become longer. When we ran the methods on the actual trees (Type I pairs of trees), both of them returned single solutions. A plausible conclusion is that as the amount of statistical error in the inferred trees increases, so does the number of maximally parsimonious solutions. The reason for this is that for shorter sequence lengths, the accuracy of the trees is poorer, i.e., they have more wrong edges. These wrong edges give an indication of more HGT events. This indication, though false, leads to larger numbers of solutions since more reconciliations become possible. Notice that the trends of the curves in Figure 8 are similar to those of the corresponding curves in Figure 3. This reflects the correlation between the number of HGT events in a minimal solution and the number of such minimal solutions, as stated in Theorem 1. However, at the same time, the proof of Theorem 1 illustrates cases where multiple solutions may exist (e.g., Figure 7 illustrates three "equivalent" HGT edges, that arise under a very specific case of incongruence between the species and gene trees). This indicates, that in some cases, even though the number of HGT events is smaller for case X than for case Y, it may be that the number of solutions for case X is larger than for case Y. This is the reason, for example, why the trend of the curve for the number of solutions in the case of three HGTs in Figure 8d does not match the trend of the number of HGT evens in the same case in Figure 3d. To illustrate this point further, the number of minimal HGT scenarios for the pair of trees in Figure 5 is 3^k. On the other hand, it is straightforward to devise an example of a pair of trees, where the minimum number of HGT edges required to reconcile them is $k + 1$, and the number of solutions is 1. In this scenario, the case of fewer HGT edges gives rise to exponentially more minimal solutions than the case of more HGT edges.

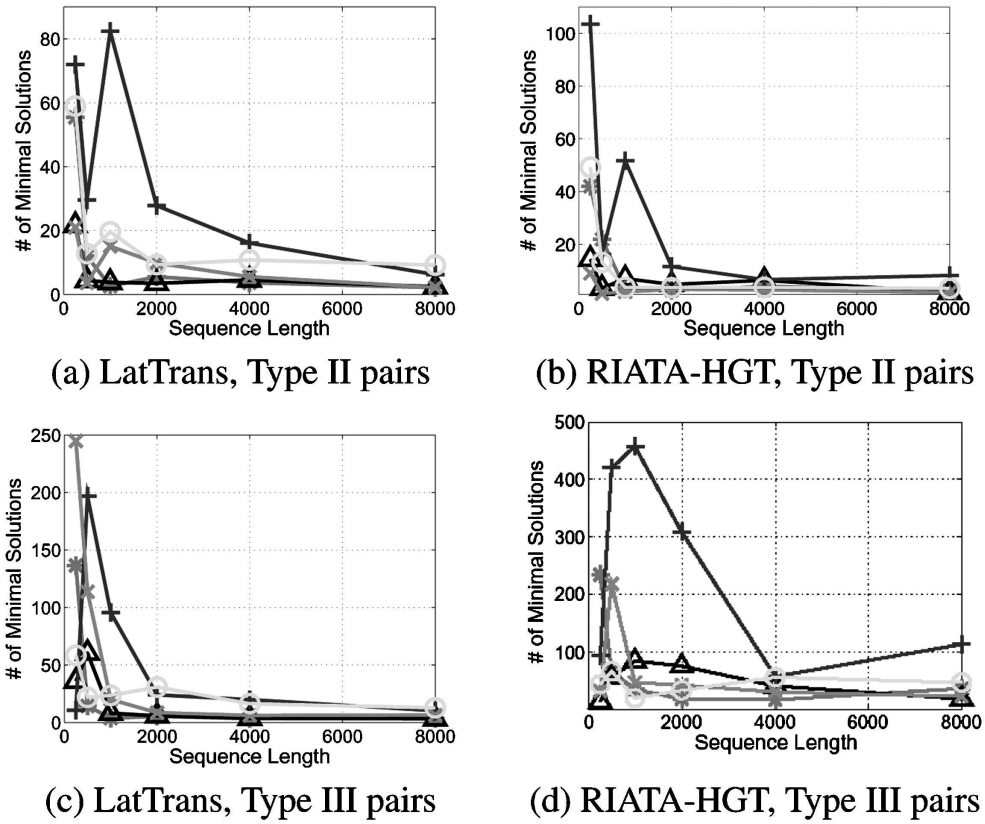

FIG. 8. The number of minimal HGT scenarios inferred by LatTrans and RIATA-HGT as a function of the sequence length. Each curve corresponds to one of the five actual numbers of HGT events: $\star$, one HGT; $\triangle$, two HGTs; $+$, three HGTs; $\times$, four HGTs; and $\circ$, five HGTs. 25-taxon trees inferred using NJ.

An important conclusion is that, in the absence of an evolutionary model of HGT, phylogeny-based HGT detection methods should be designed to compute "all" possible solutions. As illustrated in Figure 5, the number of such solutions may be exponential, though. A measure that assigns support to these solutions is imperative, so that they can be rank ordered.

5. INCORPORATING HGT INTO THE COALESCENT

As we described in Section 2, phylogenetic incongruence may occur due to various processes, of which HGT is only one. Another such process is lineage sorting, whose effect and confusing signal to HGT detection is particularly important when analyzing genes of closely related organisms. In this section, we augment the coalescent model by incorporating HGT, thus providing a framework for stochastically distinguishing among these two processes as the actual source of phylogenetic incongruence.

Lineage sorting occurs because of random contribution of each individual to the next generation. Some fail to have offsprings while some happen to have multiple offsprings. In population genetics, this process was first modeled by R.A. Fisher and S. Wright, in which each gene of the population at a particular generation is chosen independently from the gene pool of the previous generation, regardless of whether the genes are in the same individual or in different individuals. Under the Wright-Fisher model, "the coalescent" considers the process backward in time (Kingman, 1982; Hudson, 1983b; Tajima, 1983). That is, the ancestral lineages of genes of interest are traced from offsprings to parents. A coalescent event occurs when two (or sometimes more) genes are originated from the same parent, which is called the most recent common ancestor (MRCA) of the two genes.

The basic process can be treated as follows. Consider a pair of genes at time τ_1 in a random mating haploid population. The population size at time τ is denoted by $N(\tau)$. The probability that both genes are from the same parental gene at the previous generation (time $\tau_1 + 1$) is $1/N(\tau_1 + 1)$. Therefore, starting at τ_1, the probability that the coalescence between the pair occurs at τ_2 is given by

$$Prob(\tau_2) = \frac{1}{N(\tau_2)} \prod_{\tau=\tau_1+1}^{\tau_2-1} \left(1 - \frac{1}{N(\tau)}\right). \quad (1)$$

When $N(\tau)$ is constant, the probability density distribution (pdf) of the coalescent time (i.e., $t = \tau_2 - \tau_1$) is given by a geometric distribution, and can be approximated by an exponential distribution for a large N:

$$Prob(t) = \frac{1}{N} e^{-t/N}. \quad (2)$$

The coalescent process is usually ignored in phylogenetic analysis, but has a significant effect (causing lineage sorting) when closely related species are considered (Hudson, 1983a; Takahata, 1989; Rosenberg, 2002). The situation of Figure 1b is reconsidered under the framework of the coalescent in Figure 9. Here, it is assumed that species A and B split $T_1 = 5$ generations ago, and the ancestral species of A and B and species C split $T_2 = 19$ generation ago. The ancestral lineage of a gene from species A and that from B meet in their ancestral population at time $\tau = 6$, and they coalesce at $\tau = 33$, which predates T_2, the speciation time between (A, B) and C. The ancestral lineage of B enters in the ancestral population of the three species at time $\tau = 20$, and first coalesces with the lineage of C. Therefore, the gene tree is represented by $A(BC)$ while the species tree is $(AB)C$. That is, the gene tree and species tree are "incongruent." Under the model in Figure 9, the probability that the gene tree is congruent with the species tree is 0.863, which is one minus the product of the probability that the ancestral lineages of A and B do not coalesce between $\tau = 6$ and $\tau = 9$, and the probability that the first coalescence in the

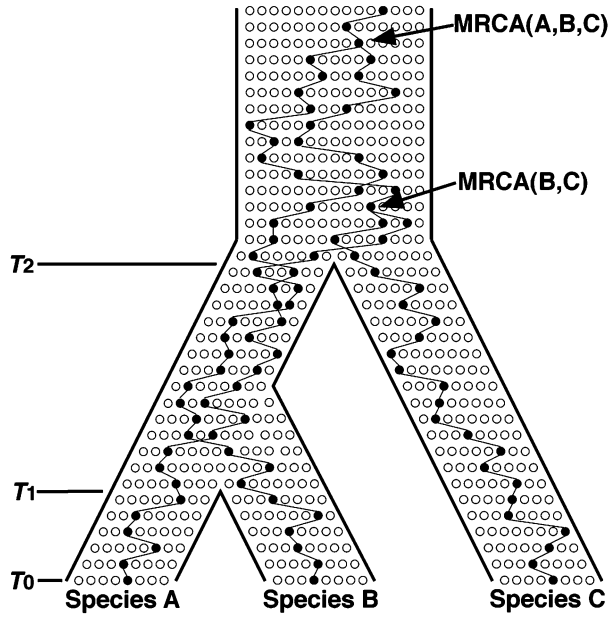

FIG. 9. An illustration of the coalescent process in a three species model with discrete generations. The process is considered backward in time from present, T_0, to past. Circles represent haploid individuals. We are interested in the gene tree of the three genes (haploids) from the three species. Their ancestral lineages are represented by closed circles connected by lines. A coalescent event occurs when a pair of lineages happen to share a single parental gene (haploid).

ancestral population of the three species occurs between (A and C) or (B and C). The former probability is $\frac{15}{16}\frac{14}{15}\frac{12}{13}\frac{11}{12}\frac{10}{11}\frac{8}{9}\cdots(1-\frac{7}{8})^8 = 0.26$ and the latter is $\frac{2}{3}$.

Under the three-species model (Fig. 9), there are three possible types of gene tree, $(AB)C$, $(AC)B$ and $A(BC)$. Let $Prob$[(AB)C], $Prob$[(AC)B] and $Prob$[A(BC)] be the probabilities of the three types of gene tree. These three probabilities are simply expressed with a continuous time approximation when all populations have equal and constant population sizes, N, where N is large:

$$Prob[(\mathrm{AB})\mathrm{C}] = 1 - \frac{2}{3}e^{-(T_2-T_1)/N}, \tag{3}$$

and

$$Prob[(\mathrm{AC})\mathrm{B}] = Prob[\mathrm{A}(\mathrm{BC})] = \frac{1}{3}e^{-(T_2-T_1)/N}. \tag{4}$$

Figure 10a shows the three probabilities as functions of $(T_2 - T_1)/N$.

It is important to notice that the estimation of the gene tree from DNA sequence data is based on the nucleotide differences between sequences, and that the gene tree is sometimes unresolved. One of the reasons for that is a lack of nucleotide differences such that DNA sequence data are not informative enough to resolve the gene tree. This possibility strongly depends on the mutation rate. Let μ be the mutation rate per region per generation, and consider the effect of mutation on the estimation of the gene tree. We consider the simplest model of mutations on DNA sequences, the infinite site model (Kimura, 1969), in which mutation rate per site is so small that no multiple mutations at a single site are allowed. Consider a gene tree, $(AB)C$, and suppose that we have a reasonable outgroup sequence such that we know the sequence of the MRCA of the three sequences. It is obvious that mutations on the internal branch between the MRCA of the three and the MRCA of A and B are informative. If at least one mutation occurred on this branch, the gene tree can be resolved from the DNA sequence alignment. This effect is investigated by assuming that the number of mutations on a branch with length t follows a Poisson distribution with mean μt. Figure 10b shows the probability that the gene tree is resolved; $T_2 - T_1 = 0.5N$ generations is assumed so that the probability that the gene tree is $(AB)C$ is about 0.6. As expected, as the mutation rate increases, the probability that the gene tree is resolved from the sequence alignment increases, and this probability exceeds 90% when $N\mu > 1.52$. Similar results are obtained for the other two types of trees, $(AC)B$ and $A(BC)$, that appear with probability 0.2 for each (Fig. 10b).

Thus far, we have shown that the gene tree is not always identical to the species tree even considering vertical evolution. With keeping this in mind, let us consider the effect of horizontal gene transfer (HGT)

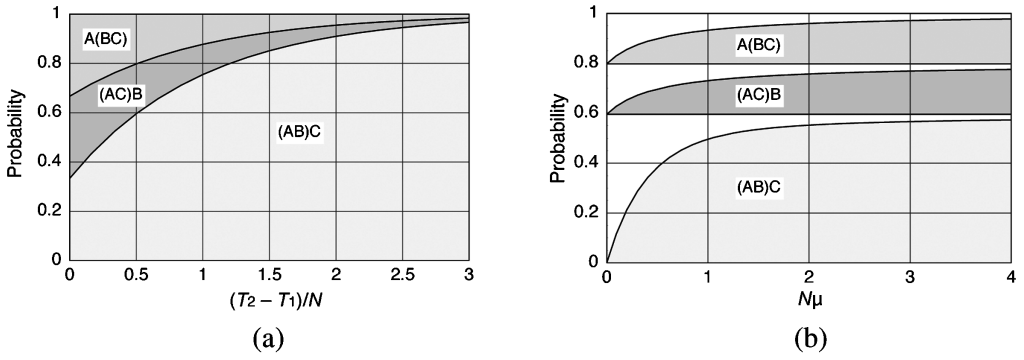

FIG. 10. **(a)** The probabilities of the three types of gene tree, (AB)C, (AC)B, and A(BC), as functions of $(T_2-T_1)/N$. **(b)** The probabilities that the gene tree is resolved from DNA sequence data. The probabilities are given as functions of the mutation rate for the three types of tree, (AB)C, (AC)B, and A(BC), when $(T_2 - T_1)/N = 0.5$. The white regions represent the probabilities that the gene tree is not resolved.

on the gene tree under the framework of the coalescent. The application of the coalescent theory to bacteria is straightforward. Rather than the Wright-Fisher model, bacterial evolution may be better described by the Moran model, which handles overlapping generations well. Suppose that each haploid individual in a bacterial population with size N has a lifespan that follows an exponential distribution with mean l. When an individual dies, another individual randomly chosen from the population replaces it to keep the population size constant. In other words, one of the $N-1$ alive lineages is duplicated to replace the dead one. Under the Moran model, the ancestral lineages of individuals of interest can be traced backward in time, and the coalescent time between a pair of individuals follows an exponential distribution with mean $lN/2$ (Ewens, 1979; Rosenberg, 2005). This means that one half of the mean lifetime in the Moran model corresponds to one generation in the Wright-Fisher model. It may usually be thought that HGT can be detected when the gene tree and species tree are incongruent (see Section 2). However, the situation is complicated when lineage sorting is also involved. Consider a model with three species, A, B, and C, in which an HGT event occurs from species B to C. Suppose the ancient circular genome has a single copy of a gene as illustrated in Figure 11a. Let a, b and c be the focal orthologous genes in the three species, respectively. At time T_h, a gene escaped from species B and was inserted in a genome in species C at T_i, which is denoted by c'. Since HGT is assumed to be instantaneous at the scale of evolution, in reality, it is always the case that $T_i = T_h$. However, since these times are estimated in practice, it may be the case that $T_h < T_i$. For example, if a gene duplication occurs in lineage b in Figure 11a, and one of the two in-paralogs is transferred to c, then the estimated time T_h would be the duplication time, which is earlier than the actual time of the HGT events, T_i.

Following the HGT event, c was physically deleted from the genome, so that each of the three species currently has a single copy of the focal gene. If there is no lineage sorting, the gene tree should be $a(bc')$. Since this tree is incongruent with the species tree, $(AB)C$, we could consider it as an evidence for HGT. However, as shown in Section 2, lineage sorting could also produce the incongruence between the gene tree and species tree without HGT. It is also important to note that lineage sorting, coupled with HGT, could produce a congruent gene tree, as illustrated in Figure 11a. Although b and c' have a higher chance to coalesce first, the probability that the first coalescence occurs between a and b or between a and c' may not be negligible especially when $T_1 - T_h$ is short. The probabilities of the three types of gene tree can be formulated under this tri-species model with HGT as illustrated in Figure 11a. Here, T_h could exceed T_1, in such a case it can be considered that HGT occurred before the speciation between A and B. Assuming that all populations have equal (constant) population sizes, N, the three probabilities can be obtained modifying (3) and (4):

$$Prob[(\mathrm{AB})\mathrm{C}] = \begin{cases} \frac{1}{3}e^{-(T_1-T_h)/N} & \text{if } T_h \le T_1 \\ 1-\frac{2}{3}e^{-(T_h-T_1)/N} & \text{if } T_h > T_1 \end{cases}, \tag{5}$$

$$Prob[(\mathrm{AC})\mathrm{B}] = \begin{cases} \frac{1}{3}e^{-(T_1-T_h)/N} & \text{if } T_h \le T_1 \\ \frac{1}{3}e^{-(T_h-T_1)/N} & \text{if } T_h > T_1 \end{cases}, \tag{6}$$

and

$$Prob[\mathrm{A}(\mathrm{BC})] = \begin{cases} 1-\frac{2}{3}e^{-(T_1-T_h)/N} & \text{if } T_h \le T_1 \\ \frac{1}{3}e^{-(T_h-T_1)/N} & \text{if } T_h > T_1 \end{cases}. \tag{7}$$

Figure 11b shows the three probabilities assuming $T_1 = 2N$ and $T_2 = 3N$.

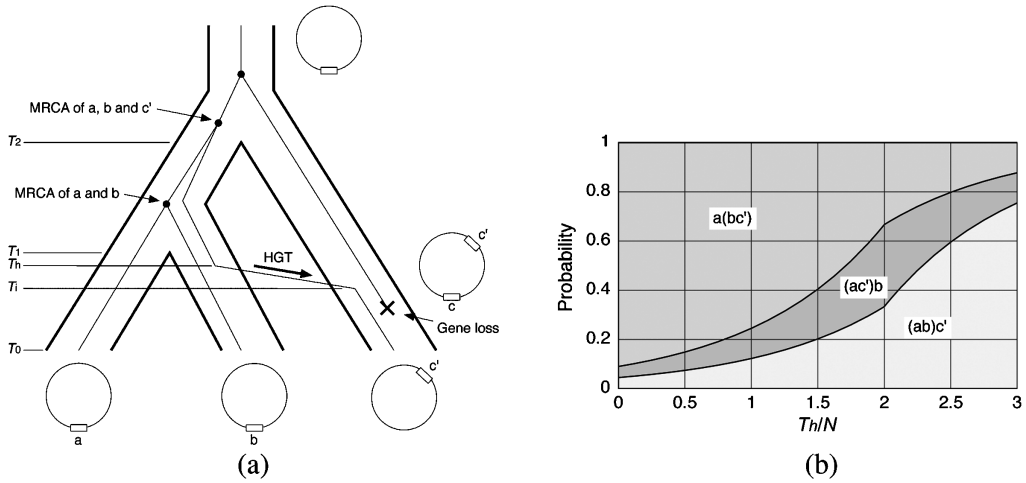

FIG. 11. **(a)** A three bacterial species model with an HGT event. A demonstration that a congruent tree could be observed even with HGT. **(b)** The probabilities of the three types of gene tree, $(ab)c'$, $(ac')b$, and $a(bc')$, as functions of T_h/N. $T_1 = 2N$ and $T_2 = 3N$ are assumed.

5.1. Enumerating valid coalescent scenarios

As mentioned above, a coalescent event occurs when two (or sometimes more) genes are originated from the same parent, which is called the most recent common ancestor (MRCA) of the two genes. A *valid coalescent scenario* for a gene tree GT and a species tree ST is a list of coalescence events in the gene tree together with the edges of the species tree on which they occur (Degnan and Salter, 2005; Rosenberg, 2007). For example, I and F in the gene tree of Figure 12 cannot coalesce at edge 2 in the species tree, and hence any coalescent scenario in which I and F coalesce at that edge is invalid. On the other hand, there are valid coalescent scenarios in which all taxa coalesce at edge 3. Rosenberg (2007) has recently provided a closed-form formula, rather than an algorithm, for computing the number of all valid coalescent scenarios of a gene within the branches of a species tree. We now provide a polynomial-time dynamic programming algorithm for enumerating all valid coalescent scenarios, given a pair of species/gene trees, over the same set of taxa, but not necessarily bifurcating or having the same topology. The divide-and-conquer nature of our algorithm allows for computing sharing among the many (potentially exponential) valid coalescent scenarios, and hence for more efficient implementation to list these scenarios.

We first start with some definitions and terminology.

Let T be a rooted tree leaf-labeled by a set $\mathcal{X}$ of taxa; i.e., there is a bijection $f : L(T) \to \mathcal{X}$, where $L(T)$ denotes the set of the tree leaves. The set $E(T)$ denotes the set of all internal edges (including an edge incoming into the root, which we denote by *re*). Further, the tree edges are labeled via a post-order numbering; i.e., there is a bijection $h^T : E(T) \to \{1, \ldots, n-1\}$, where $n = |\mathcal{X}|$, and h^T respects a

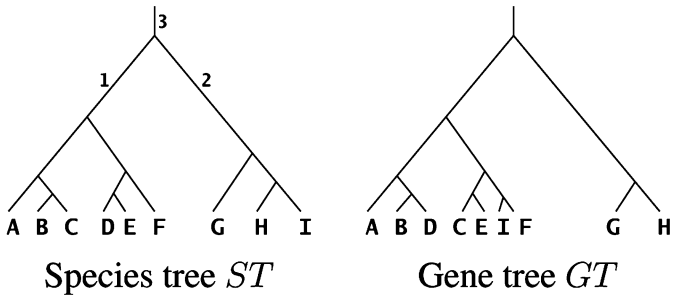

FIG. 12. Illustration of incongruent species/gene trees. The three numbered edges in tree ST are the only elements of the set $\mathcal{E}$, which is the set of all species tree edges on which at least one cluster of taxa from the gene tree can coalesce at.

post-order labeling. Tree T induces a set $C_T = \{c_e^T : c_e^T \subseteq \mathcal{X},\ e \in E(T)\}$ of *clusters* of taxa, where c_e^T is the set of all leaves in $\mathcal{X}$ which are "under" edge e in T. The topology of the tree T naturally defines a partial order $\subseteq_T$ on C (indeed, the topology of T is the Hasse diagram of this partial order).

Definition 2. *Given two trees, ST and GT, over the same set $\mathcal{X}$ of taxa, a (valid) coalescent history is a mapping $\alpha : C_{GT} \to E(ST)$, such that:*

1. *For every $X \in C_{GT}$, $X \subseteq c^{ST}_{\alpha(X)}$, and*
2. *For every two edges $e_1, e_2 \in E(GT)$, if $h^{GT}(e_1) < h^{GT}(e_2)$ then $h^{ST}(\alpha(e_1)) \leq h^{ST}(\alpha(e_2))$.*

Condition (1) of Definition 2 states that a cluster X of taxa can coalesce on branches of the species tree that are "above" the most recent common ancestor (MRCA) of X in the tree. Condition (2) of Definition 2 states that the coalescent events of the gene within the branches of the species tree must respect the gene tree topology.

Let *ST* and *GT* be the species and gene trees, respectively. Let the edges of *ST* be numbered as above, and let C be the set of all clusters (of size ≥ 2) of taxa in the tree *GT*. We write $v = ca_{ST}(x)$, for $x \in C$, to denote that v is a common ancestor (node) of the set x of taxa in tree *ST*. We write InEdge(v) to denote the edge incident into node v (in v's tree). The set of all edges in *ST* on which any cluster in C can coalesce at is $\mathcal{E} = \{\text{InEdge}(v) : v = ca_{ST}(x),\ x \in C\}$. For example, $\mathcal{E} = \{1, 2, 3\}$ for the species and gene trees in Figure 12. We say that $e \in \mathcal{E}$ is a "lowest" edges if there does not exist any other edge $e' \in \mathcal{E}$ such that e lies on the path from the root of the tree to e'.

We define *children* of a cluster $c \in C$ as $\text{Children}(c) = \{c' \in C : c' \subset c,\ \text{and}\ \nexists c'' \in C \text{ s.t. } [c' \subset c'' \wedge c'' \subset c]\}$. Notice that Children($c$) induces a partition of c. In other words, for every $c_1, c_2 \in C$, $c_1 \neq \emptyset$, $c_1 \cap c_2 = \emptyset$, and $\cup_{c' \in C} c' = c$.

With every cluster $c \in C$, we associate the set p_c, which is $p_c = \{\text{InEdge}(v) : v = ca_{ST}(c)\}$. In other words, p_c is actually the path of edges in the species tree on which the taxa of cluster c could coalesce at. For example, for cluster $c = ABD$ in Figure 12 we have $p_c = \{1, 3\}$. For every cluster c and edge e, we define $\rho_c(e)$ to be the total number of all valid coalescence scenarios of leaves in c when the MRCA of all leaves in c lies on edge e. The recursive algorithm, **Compute**ρ, outlined in Figure 13, computes the values of $\rho_c(e)$ for a cluster $c \in C$ and edge $e \in \mathcal{E}$. The recursion can be eliminated in a straightforward manner if the computation of ρ is done in a bottom-up fashion.

The number of valid coalescent scenarios, given a species tree *ST* and a gene tree *GT* is **Compute**ρ(*ST*, *GT*, L(*ST*), *re*). We illustrate the algorithm on the trees shown in Figure 12. The results are shown in Table 1.

The following two theorems establish the correctness and running time of computing the number of valid coalescent scenarios of a pair of species and gene trees.

Compute$\rho(ST, GT, c, e)$

1. **If** $e \notin p_c$, **then** $\rho_c(e) = 0$.
2. **If** c is a "lowest" cluster (i.e., there is no cluster $c' \in C$ such that $c' \subset c$) or e is a lowest edge in $\mathcal{E}$, **then** $\rho_c(e) = 1$.
3. **If** c and e do not satisfy the conditions of 1 and 2, **then**

$$\rho_c(e) = \prod_{c' \in \text{Children}(c)} \left[\sum_{(e' \in p_{c'} \text{ and } e' \leq e)} \textbf{Compute}\rho(ST, GT, c', e') \right].$$

4. **Return** $\rho_c(e)$.

FIG. 13. Algorithm **Compute**ρ for computing the value $\rho_c(e)$ for a cluster $c \in C$ and an edge $e \in \mathcal{E}$, given a species tree *ST* and a gene tree *GT*.

TABLE 1. CLUSTERS IN C, AND THE VALUES OF p_c AND $\rho_c(e)$ FOR THE TREES IN FIGURE 12

Cluster c	*Set* p_c	$\rho_c(e)\ \forall e \in \mathcal{E}$
1: BD	1,3	$\rho_1(1) = \rho_1(3) = 1;\ \rho_1(2) = 0$
2: ABD	1,3	$\rho_2(1) = 1;\ \rho_2(2) = 0;\ \rho_2(3) = \rho_1(1) + \rho_1(3) = 2$
3: CE	1,3	$\rho_3(1) = \rho_3(3) = 1;\ \rho_3(2) = 0$
4: IF	3	$\rho_4(1) = \rho_4(2) = 0;\ \rho_4(3) = 1$
5: CEIF	3	$\rho_5(1) = \rho_5(2) = 0;\ \rho_5(3) = (\rho_3(1) + \rho_3(3))(\rho_4(3)) = 2 \cdot 1 = 2$
6: ABDCEIF	3	$\rho_6(1) = \rho_6(2) = 0;\ \rho_6(3) = (\rho_2(1) + \rho_2(3))(\rho_5(3)) = 3 \cdot 2 = 6$
7: GH	2,3	$\rho_7(1) = 0;\ \rho_7(2) = \rho_7(3) = 1;$
8: ABCDEFGHI	3	$\rho_8(1) = \rho_8(2) = 0;\ \rho_8(3) = (\rho_6(3))(\rho_7(2) + \rho_7(3)) = 6 \cdot 2 = 12$

The total number of valid coalescent scenarios is 12, which is $\rho_8(3)$.

Theorem 2. **Compute**ρ(*ST*, *GT*, L(*ST*), *re*) *is the number of valid coalescent scenarios of the species tree ST and the gene tree GT.*

When the recursion in algorithm **Compute**ρ is replaced by bottom-up computation, the algorithm takes $O(n^2)$ time, since for each cluster c, the algorithm considers only its children (whose ρ values have already been computed). Given that there are $O(n)$ clusters in the gene tree (each defined by an internal edge in the tree), and that for each cluster there may be $O(n)$ children, the computation can be achieved in $O(n^2)$ time.

Theorem 3. **Compute**ρ(*ST*, *GT*, L(*ST*), *re*) *can be computed in* $O(n^2)$ *time, where* $n = |L(ST)| = |L(GT)|$.

5.2. Coalescent versus HGT scenarios

Experimental settings. In this experiment we compared the number of valid coalescent and minimal HGT scenarios returned by the algorithm described in Section 5.1 and the extended RIATA-HGT heuristic, respectively, for a range of numbers and diameters of simulated incongruence events.[2] The diameter of an incongruence event is the number of tree edges separating the two endpoints of the SPR move to which it corresponds. This quantity reflects the "locality" of the incongruities between the two trees. For each number $n \in \{1, 2, 3, 4\}$ and diameter $d \in \{2, 4, 6, 8\}$ of simulated incongruence events, we generated twenty pairs $\langle t, t' \rangle$ of 20-taxon trees, where t was generated using the `r8s` tool (Sanderson, n.d.), and t' was obtained from t by simulating n random incongruence events, each of diameter d. The coalescent algorithm and extended RIATA-HGT were run on each such pair within the data set and the numbers of solutions computed were averaged to give a single data point for each method and combination of diameter/number of incongruence events. The results of this analysis are shown in Figure 14.

Results and discussion. Figure 14 clearly shows that there is a correlation between the diameter of incongruence events and the number of valid coalescent and minimal HGT scenarios. Small diameters reflect that the incongruence occurs between two very close taxa, whereas large diameters reflect incongruence between two very distant taxa. As the diameter gets larger, the number of edges between the MRCA of taxa and the root becomes smaller, and hence we would expect the number of valid coalescent scenarios to become smaller. And this is exactly what we see in the figure. For diameter of 2, the number of such scenarios is over 200 million. However, there is a sharp decrease in that number as the diameter increases. On the other hand, even though we see a similar trend in the decrease of number of HGT scenarios as the diameter increases, the actual number of minimal HGT scenarios is drastically much smaller. Even for the

[2]We simulated an incongruence event by a *subtree prune and regraft* (SPR) move, where these moves were added with certain restrictions, as described in Section 3.

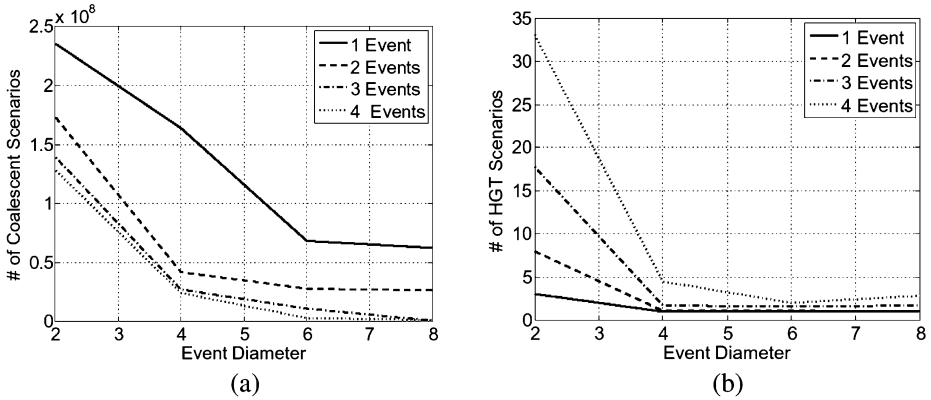

FIG. 14. The numbers of valid coalescent scenarios **(a)** and minimal HGT scenarios **(b)** as a function of the diameter of the incongruence events. Each curve corresponds to datasets with one, two, three, or four incongruence events.

smallest diameter of 2, the number of minimal HGT scenarios is about 33, and as the diameter increases, the number of solutions converges to 1.

A significant observation from Figure 14b is that we either have small diameter and large number of solutions, or a large diameter and small number of solutions. This observation proves crucial to the improvements achieved by the algorithmic techniques described in the previous section, since small diameters indicate more pairs in the decomposition, and hence more efficiency, and small number of solutions for large diameters indicate very small equivalence classes in large components of the decomposition.

6. CONCLUSION

In this paper, we showed that error in inferred trees has a negative impact on the estimates made by phylogeny-based HGT detection methods. These results provide a set of conclusions. First, to obtain accurate estimates of HGT based on tree incongruence, poorly supported edges of reconstructed trees should be removed. Though an important task to conduct, removing (or contracting) poorly supported edges is not a straightforward task, since standard methods that are in common use for estimating branch support, such as bootstrapping and posterior probabilities in Bayesian analyses, have been shown to be overly "conservative" or "liberal" under various circumstances (Ruths and Nakhleh, 2006). Second, more than one maximally parsimonious solution (a solution that has the minimum number of HGT edges, or events, to explain the species and gene tree incongruence) may exist, and hence HGT detection methods should search for all such solutions. In this preliminary work, we have studied the effect of error in inferred trees on the accuracy of HGT detection methods, both in terms of the minimum number of events computed as well as the number of such minimal solutions. One of our immediate goals is to study the performance of these methods in terms of the locations (donors and recipients) of inferred HGT; for this task, we will use the distance measures proposed in Moret et al. (2004). Further, we will study the effects of the aforementioned factors, using both simulated and biological data, on the performance of several currently available tools for HGT detection (Hallett and Lagergren, 2001; Makarenkov, 2001; Addario-Berry et al., 2003; Nakhleh et al., 2005; MacLeod et al., 2005; Beiko and Hamilton, 2006).

Further, lineage sorting due to the coalescent process acts as a noise for detecting and reconstructing HGT based on tree incongruence, sometimes mimicking the evidence for HGT and sometimes concealing evidence of HGT. Therefore, to distinguish HGT and lineage sorting, a stochastic framework based on the theory introduced in Section 5 is needed. We only considered very simple cases with three species here, and we will extend the theory to more general cases.

Finally, we designed a polynomial-time dynamic programming algorithm for enumerating all valid coalescent scenarios that reconcile a pair of species and gene trees. This algorithm may be used as a core component in statistical methods for reconciling species and gene trees.

ACKNOWLEDGMENTS

We are grateful to Noah Rosenberg for comments and discussion of the problem of enumerating valid coalescent scenarios. We would like to thank the three anonymous reviewers for their valuable comments.

This work is supported in part by the Department of Energy grant DE-FG02-06ER25734 (Luay Nakhleh), the National Science Foundation grant CCF-0622037 (Hideki Innan and Luay Nakhleh), and the George R. Brown School of Engineering Roy E. Campbell Faculty Development Award (Luay Nakhleh).

REFERENCES

Addario-Berry, L., Hallett, M.T., and Lagergren, J. 2003. Towards identifying lateral gene transfer events. *Proc. 8th Pacific Symp. Biocomput.*, 279–290.

Atteson, K. 1999. The performance of the neighbor-joining methods of phylogenetic reconstruction. *Algorithmica* 25, 251–278.

Beiko, R.G., and Hamilton, N. 2006. Phylogenetic identification of lateral genetic transfer events. *BMC Evol. Biol.* 6, 15–31.

Bordewich, M., and Semple, C. 2005. On the computational complexity of the rooted subtree prune and regraft distance. *Ann. Combin.* 1–15.

Daubin, V., Moran, N.A., and Ochman, H. 2003. Phylogenetics and the cohesion of bacterial genomes. *Science* 301, 829–832.

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

Doolittle, W.F., Boucher, Y., Nesbo, C.L., et al. 2003. How big is the iceberg of which organellar genes in nuclear genomes are but the tip? *Phil. Trans. R. Soc. Lond. B. Biol. Sci.* 358, 39–57.

Ewens, W.J. 1979. *Mathematical Population Genetics*. Springer-Verlag, Berlin.

Hallett, M.T., and Lagergren, J. 2001. Efficient algorithms for lateral gene transfer problems. *Proc. RECOMB 2001*, 149–156.

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

Hudson, R.R. 1983b. Properties of the neutral allele model with intergenic recombination. *Theor. Popul. Biol.* 23, 183–201.

Jin, G., Nakhleh, L., and Than, C. 2007. Integrating sequence and topology for efficient and accurate detection of horizontal gene transfer (submitted).

Kimura, M. 1969. The number of heterozygous nucleotide sites maintained in a finite population due to steady flux of mutations. *Genetics* 61, 893–903.

Kingman, J.F.C. 1982. The coalescent. *Stochast. Proc. Appl.* 13, 235–248.

Kunin, V., and Ouzounis, C.A. 2003. The balance of driving forces during genome evolution in prokaryotes. *Genome Res.* 13, 1589–1594.

Kunin, V., Goldovsky, L., Darzentas, N., et al. 2005. The net of life: reconstructing the microbial phylogenetic network. *Genome Res.* 15, 954–959.

Lerat, E., Daubin, V., and Moran, N.A. 2003. From gene trees to organismal phylogeny in prokaryotes: the case of the γ-proteobacteria. *PLoS Biol.* 1, 1–9.

MacLeod, D., Charlebois, R.L., Doolittle, F., et al. 2005. Deduction of probable events of lateral gene transfer through comparison of phylogenetic trees by recursive consolidation and rearrangement. *BMC Evol. Biol.* 5, 27–37.

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

Makarenkov, V. 2001. T-REX: reconstructing and visualizing phylogenetic trees and reticulation networks. *Bioinformatics* 17, 664–668.

Moret, B.M.E., Nakhleh, L., Warnow, T., et al. 2004. Phylogenetic networks: modeling, reconstructibility, and accuracy. *IEEE/ACM Trans. Comput. Biol. Bioinform.* 1, 13–23.

Nakhleh, L., Warnow, T., and Linder, C.R. 2004. Reconstructing reticulate evolution in species—theory and practice. *RECOMB 2004*, 337–346.

Nakhleh, L., Ruths, D., and Wang, L.S. 2005. RIATA-HGT: a fast and accurate heuristic for reconstrucing horizontal gene transfer. *Lect. Notes Comput. Sci.* 3595, 84–93.

Ochman, H., Lawrence, J.G., and Groisman, E.A. 2000. Lateral gene transfer and the nature of bacterial innovation. *Nature* 405, 299–304.

Paulsen, I.T., Banerjei, L., Myers, G.S., et al. 2003. Role of mobile DNA in the evolution of Vancomycin-resistant *Enterococcus faecalis*. *Science* 299, 2071–2074.

Rambaut, A., and Grassly, N.C. 1997. Seq-gen: an application for the Monte Carlo simulation of DNA sequence evolution along phylogenetic trees. *Comput. Appl. Biosci.* 13, 235–238.

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

Rosenberg, N.A. 2005. Gene genealogies. In: Fox, C.W., and Wolf, J.B., eds., *Evolutionary Genetics: Concepts and Case Studies*. Chapter 15. Oxford University Press, New York.

Rosenberg, N.A. 2007. A recursion for the number of valid coalescent histories. *J. Comput. Biol.* (in press).

Ruths, D., and Nakhleh, L. 2006. Techniques for assessing phylogenetic branch support: a performance study. *Proc. Fourth Asia-Pacific Bioinform. Conf. (APBC 06)*, 187–196.

Saitou, N., and Nei, M. 1987. The neighbor-joining method: A new method for reconstructing phylogenetic trees. *Mol. Biol. Evol.* 4, 406–425.

Sanderson, M. n.d. `r8s` software package. Available at: *http://loco.ucdavis.edu/r8s/r8s.html*. Accessed April 1, 2007.

Swofford, D.L. 1996. *PAUP*: Phylogenetic Analysis Using Parsimony (and Other Methods)*, Version 4.0. Sinauer Associates, Sunderland, MA.

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

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

Welch, R.A., Burland, V., Plunkett, G., et al. 2002. Extensive mosaic structure revealed by the complete genome sequence of uropathogenic *Escherichia coli*. *Proc. Natl. Acad. Sci. USA* 99, 17020–17024.

Zwickl, D., and Hillis, D. 2002. Increased taxon sampling greatly reduces phylogenetic error. *Syst. Biol.* 51, 588–598.

Address reprint requests to:
Dr. Luay Nakhleh
Department of Computer Science
Rice University
6100 Main St.
Houston, TX 77005

E-mail: nakhleh@cs.rice.edu