Identifiability Issues in Phylogeny-Based Detection of Horizontal Gene Transfer

Cuong Than[1], Derek Ruths[1], Hideki Innan[2], and Luay Nakhleh[1,⋆]

[1] Dept. of Computer Science, Rice University, Houston, TX, USA
[2] Human Genetics Center, The University of Texas Health Science Center, Houston, TX, USA
nakhleh@cs.rice.edu

Abstract. Prokaryotic organisms share genetic material across species boundaries by means of a process known as *horizontal gene transfer* (HGT). Detecting this process bears great significance on understanding prokaryotic genome diversification and unraveling their complexities. Phylogeny-based detection of HGT is one of the most commonly used approaches 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. Hence, methods that adopt this approach compare gene and species trees, and infer a set of HGT events to reconcile the differences among these trees.

In this paper, we address some of the identifiability issues that face phylogeny-based detection of HGT. In particular, we show the effect of inaccuracies in the reconstructed (species and gene) trees on inferring the correct number of HGT events. Further, we show that a large number of maximally parsimonious HGT scenarios may exist. These results indicate that accurate detection of HGT requires accurate reconstruction of individual trees, and necessitates the search for more than a single scenario to explain gene tree disagreements. Finally, we show that disagreements among trees may be a result of not only HGT, but also *lineage sorting*, and make initial progress on incorporating HGT into the coalescent model, so as to stochastically distinguish between the two and make an accurate reconciliation. This contribution is very significant, particularly when analyzing closely related organisms.

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); e.g., see [4,19]. There has been a big "ideological and rhetorical" gap between the researchers believing that HGT is so rampant that a prokaryotic phylogenetic tree is useless and those who believe HGT is mere "background noise" that does not affect the reconstructibility of a phylogenetic tree for bacterial genomes.

⋆ Corresponding author.

G. Bourque and N. El-Mabrouk (Eds.): RECOMB-CG 2006, LNBI 4205, pp. 215–229, 2006.
© Springer-Verlag Berlin Heidelberg 2006

Supporting arguments for these two views have been published. For example, the heterogeneity of genome composition between closely related strains (only 40% of the genes in common with three *E. coli* strains [29]) supports the former view, whereas the well-supported phylogeny reconstructed by Lerat *et al.* from about 100 "core" genes in γ-Proteobacteria [13] 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 (e.g., [5]), is considered a primary explanation of incongruence among gene phylogenies, and is a significant obstacle to reconstructing the Tree of Life [3].

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. 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).

We draw several conclusions from this work. First, to obtain accurate estimates of HGT based on tree incongruence, poorly supported edges of reconstructed trees should be removed; this is a hard task, but is very important to conduct. 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. 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 [14]. 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 1(a), 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 1(b) 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 1(c), 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.

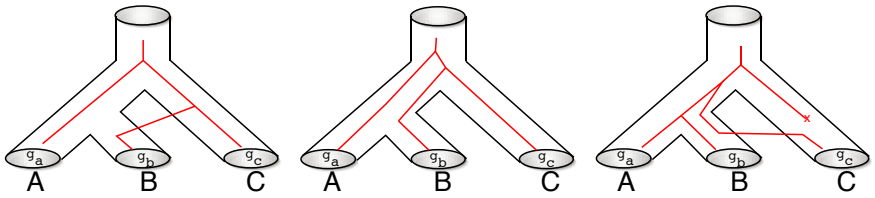


Fig. 1. (a) Gene tree that disagrees with the species tree due to (a) HGT from C to B and (b) lineage sorting due to random genetic drift. In (c), the effect of the HGT event (from B to C) is "canceled out" by random genetic drift, resulting in congruent species and gene trees.

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 [16,12]. 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; e.g., see [7,18,2,17,15].

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 Setting. We used the `r8s` tool [25] 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] 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 [20]. We used the parameter settings of [30]. Then, from each sequence alignment, we reconstructed a tree TNJ using the Neighbor Joining (NJ) method [24], and another tree using a maximum parsimony heuristic as implemented in PAUP* [26]. 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 [7,1] and RIATA-HGT [17] (we modified the latter

[1] An SPR move simulates an HGT event.

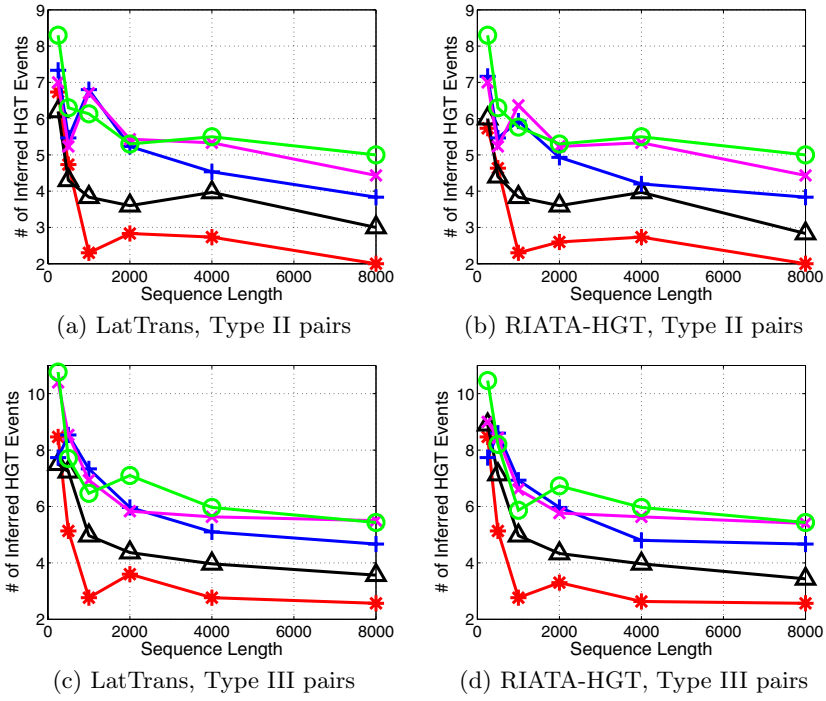
tool so that it computes multiple solutions, rather than a single solution as was originally described by the authors). Both tools were applied to three different types of pairs of trees.

Type I pairs $(T_i, T_{i,j})$**:** in this case, the species and gene trees are assumed to be correct.
Type II pairs $(T_i, TNJ^{\ell}_{i,j}[k])$ and $(T_i, TMP^{\ell}_{i,j}[k])$**:** in this case, the species tree is correct, and the gene trees are estimated (using NJ and MP, respectively).
Type III pairs $(TNJ^{\ell}_{i}[k], TNJ^{\ell}_{i,j}[k])$ and $(TMP^{\ell}_{i}[k], TMP^{\ell}_{i,j}[k])$**:** 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). 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.

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 Fig. 2. Figs. 2(a) and 2(b) 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 5 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*, 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 Figs. 2(c) and 2(d). 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. 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 sets (50- and 100-taxon trees). Therefore, it is important to



(a) LatTrans, Type II pairs (b) RIATA-HGT, Type II pairs

(c) LatTrans, Type III pairs (d) RIATA-HGT, Type III pairs

Fig. 2. 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: $\star$: 1 HGT; $\triangle$: 2 HGTs; $+$: 3 HGTs; $\times$: 4 HGTs; and $\circ$: 5 HGTs. 25-taxon trees inferred using NJ.

eliminate statistical error from trees before estimating HGT events. Ruths and Nakhleh [23] have studied the performance of various methods for eliminating wrong edges while maintaining accurate ones. This elimination, in the form of contracting poorly supported edges, leads to non-binary trees, which cannot be handled by LatTrans, although they can be handled by RIATA-HGT. Therefore, a second conclusion is that phylogeny-based HGT detection methods should be designed to handle both bi- and multi-furcating trees.

3.2 The Uniqueness of HGT Scenarios

Moret *et al.* [16] showed that a phylogenetic network that reconciles two trees need not be unique, by showing two phylogenetic networks 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 Fig. 3. 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 Fig. 4. To reconcile the differences among all k 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

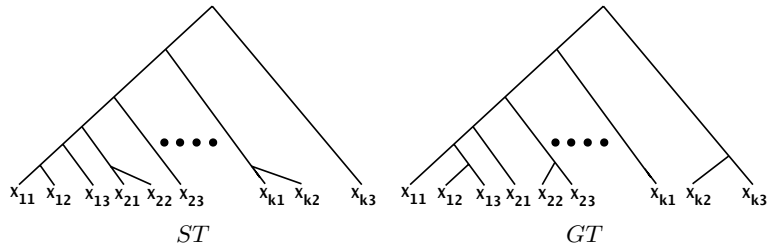


Fig. 3. A species tree ST and a gene tree GT with $3k$ leaves. The two trees differ in k places: the species tree has X_{i1} and X_{i2} as siblings, whereas the gene tree has X_{i2} and X_{i3} as siblings ($1 \leq i \leq k$). There are 3^k maximally parsimonious HGT scenarios that reconcile the two trees.

donor and recipient of a gene have to co-exist in time [16], 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 Fig. 4, 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 Fig. 3, and let us consider the triplet of taxa in Fig. 4. 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

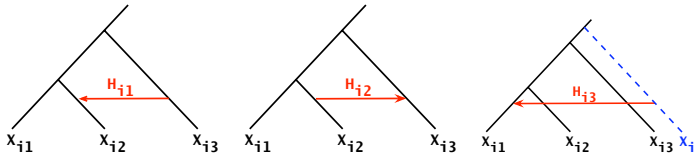


Fig. 4. The three possible scenarios for reconciling the topologies of the triplet $\langle X_{i1}, X_{i2}, X_{i3} \rangle$ based on the species and gene trees, ST and GT respectively, in Fig. 3.

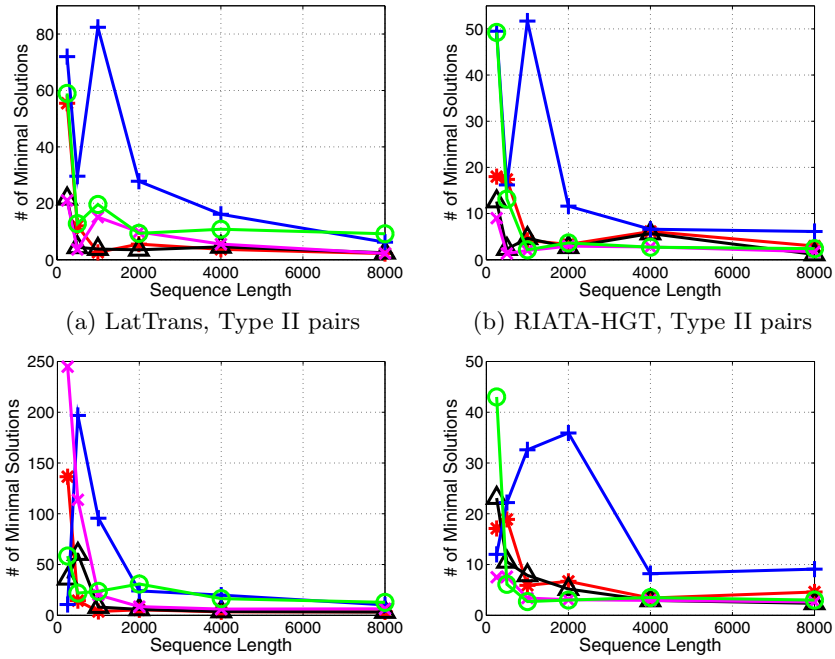
to a high degree of accuracy such that the above three cases are distinguished accurately is a very challenging task. Further, even if branch lengths are estimated accurately, if the evolutionary distance between the donor and recipient is very small, distinguishing among the cases becomes more challenging.

In our simulation study, 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 Fig. 5. 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 behind 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. The peaks around sequence lengths 1000 and 2000 in Fig. 5 coincide with the peaks in Fig. 2, which gives an indication that as the number of inferred HGT events increases, so does the number of possible solutions. An important conclusion is that phylogeny-based HGT detection methods should be designed to compute "all" possible solutions. As illustrated in Fig. 3, 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.

4 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



(a) LatTrans, Type II pairs

(b) RIATA-HGT, Type II pairs

(c) LatTrans, Type III pairs

(d) RIATA-HGT, Type III pairs

Fig. 5. 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: $\star$: 1 HGT; $\triangle$: 2 HGTs; $+$: 3 HGTs; $\times$: 4 HGTs; and $\circ$: 5 HGTs. 25-taxon trees inferred using NJ.

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 [11,9,27]. 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 the pair 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

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

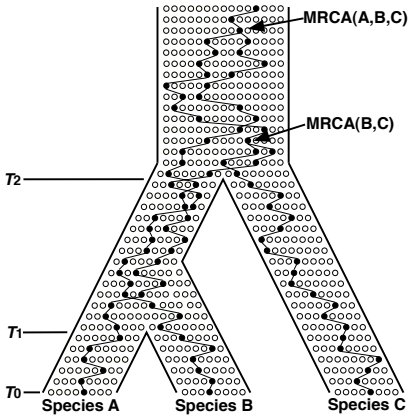
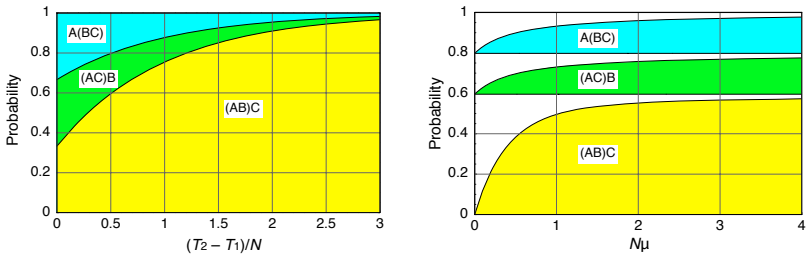


Fig. 6. An illustration of the coalescent process in a three species model with discrete generations. The process is considered backward in time from present, T_0, to past. Circles represent haploid individuals. We are interested in the gene tree of the three genes (haploids) from the three species. 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).

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:

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

The coalescent process is usually ignored in phylogenetic analysis, but has a significant effect (causing lineage sorting) when closely related species are considered [8,28,21]. The situation of Fig. 1(b) is reconsidered under the framework of the coalescent in Fig. 6. 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 $\tau = 35$, 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 Fig. 6, the probability that the gene tree is congruent with the species tree is 0.85, 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 ancestral population of the three species occur between (A and C) or (B and C). The former probability is $\frac{14}{15}\frac{12}{13}\frac{11}{12}...\frac{7}{8}\frac{7}{8} = 0.22$ and the latter is $\frac{2}{3}$.



(a) (b)

Fig. 7. (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.

Under the three-species model (Fig. 6), 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:

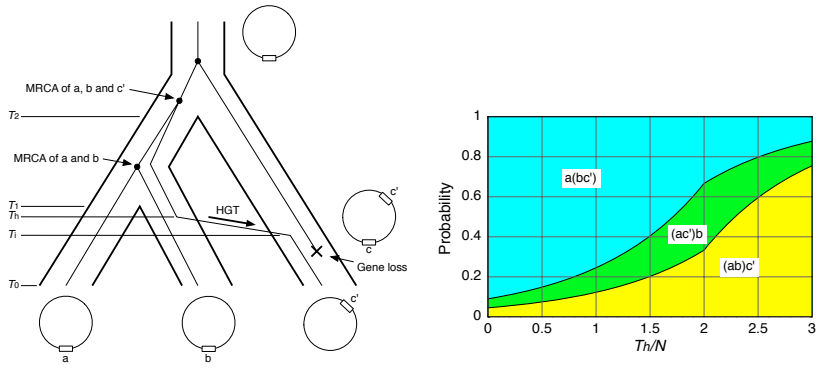
$$Prob[(\text{AB})\text{C}] = 1 - \frac{2}{3}e^{-(T_2-T_1)/N}, \tag{3}$$

and

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

Figure 7(a) 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 [10], 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. Fig. 7(b) shows the probability that the gene tree is resolved; $T_2 - T_1 = 0.5N$ generations is assumed



(a) (b)

Fig. 8. (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.

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 appears with probability 0.2 for each (see also Fig. 7(b)).

Thus far, we have shown that the gene tree is not always identical to the species tree even considering verical evolution. With keeping this in mind, let us consider the effect of horizontal gene transfer (HGT) on 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 $lN/2$ [6,22]. 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 Fig. 8(a). 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'. 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 congruent gene tree, as illustrated in Fig. 8(a). 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 Fig. 8(a). 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):

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

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

and

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

Fig. 8(b) shows the three probabilities assuming $T_1 = 2N$ and $T_2 = 3N$.

5 Conclusions and Future Work

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; this is a hard task, but is very important to conduct. 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. 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 [16].

Further, lineage sorting due to the coalescent process works as a noise for detecting and reconstructing HGT based on tree incongruence, sometimes mimicking the evidence for HGT and sometimes creating a false negative "evidence" for HGT. Therefore, to distinguish HGT and lineage sorting, a stochastic framework based on the theory introduced in Section 4 is needed. We only considered very simple cases with three species here, and we will extend the theory to more general cases.

References

1. L. Addario-Berry, M.T. Hallett, and J. Lagergren. Towards identifying lateral gene transfer events. In *Proc. 8th Pacific Symp. on Biocomputing (PSB03)*, pages 279–290, 2003.
2. M. Bordewich and C. Semple. On the computational complexity of the rooted subtree prune and regraft distance. *Annals of Combinatorics*, pages 1–15, 2005. In press.
3. V. Daubin, N.A. Moran, and H. Ochman. Phylogenetics and the cohesion of bacterial genomes. *Science*, 301:829–832, 2003.
4. W.F. Doolittle, Y. Boucher, C.L. Nesbo, C.J. Douady, J.O. Andersson, and A.J. Roger. 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, 2003.
5. I.T. Paulsen *et al.* Role of mobile DNA in the evolution of Vacomycin-resistant Enterococcus faecalis. *Science*, 299(5615):2071–2074, 2003.
6. W.J. Ewens. *Mathematical Population Genetics.* Springer-Verlag, Berlin, 1979.
7. M.T. Hallett and J. Lagergren. Efficient algorithms for lateral gene transfer problems. In *Proc. 5th Ann. Int'l Conf. Comput. Mol. Biol. (RECOMB01)*, pages 149–156, New York, 2001. ACM Press.
8. R. R. Hudson. Testing the constant-rate neutral allele model with protein sequence data. *Evolution*, 37:203–217, 1983.
9. R.R. Hudson. Properties of the neutral allele model with intergenic recombination. *Theor. Popul. Biol.*, 23:183–201, 1983.
10. M. Kimura. The number of heterozygous nucleotide sites maintained in a finite population due to steady flux of mutations. *Genetics*, 61:893–903, 1969.
11. J. F. C. Kingman. The coalescent. *Stochast. Proc. Appl.*, 13:235–248, 1982.
12. V. Kunin, L. Goldovsky, N. Darzentas, and C.A. Ouzounis. The net of life: reconstructing the microbial phylogenetic network. *Genome Research*, 15:954–959, 2005.
13. E. Lerat, V. Daubin, and N.A. Moran. From gene trees to organismal phylogeny in prokaryotes: The case of the γ-proteobacteria. *PLoS Biology*, 1(1):1–9, 2003.
14. W.P. Maddison. Gene trees in species trees. *Systematic Biology*, 46(3):523–536, 1997.
15. V. Makarenkov. T-REX: Reconstructing and visualizing phylogenetic trees and reticulation networks. *Bioinformatics*, 17(7):664–668, 2001.
16. B.M.E. Moret, L. Nakhleh, T. Warnow, C.R. Linder, A. Tholse, A. Padolina, J. Sun, and R. Timme. Phylogenetic networks: modeling, reconstructibility, and accuracy. *IEEE/ACM Transactions on Computational Biology and Bioinformatics*, 1(1):13–23, 2004.

17. L. Nakhleh, D. Ruths, and L.S. Wang. RIATA-HGT: A fast and accurate heuristic for reconstrucing horizontal gene transfer. In L. Wang, editor, *Proceedings of the Eleventh International Computing and Combinatorics Conference (COCOON 05)*, pages 84–93, 2005. LNCS #3595.
18. L. Nakhleh, T. Warnow, and C.R. Linder. Reconstructing reticulate evolution in species–theory and practice. In *Proc. 8th Ann. Int'l Conf. Comput. Mol. Biol. (RECOMB04)*, pages 337–346, 2004.
19. H. Ochman, J.G. Lawrence, and E.A. Groisman. Lateral gene transfer and the nature of bacterial innovation. *Nature*, 405(6784):299–304, 2000.
20. A. Rambaut and N. C. Grassly. Seq-gen: An application for the Monte Carlo simulation of DNA sequence evolution along phylogenetic trees. *Comp. Appl. Biosci.*, 13:235–238, 1997.
21. N. Rosenberg. The probability of topological concordance of gene trees and species tree. *Theoretical Population Biology*, 61:225–247, 2002.
22. N.A. Rosenberg. Gene genealogies. In C.W. Fox and J. B. Wolf, editors, *Evolutionary Genetics: Concepts and Case Studies*, chapter 15. Oxford Univ. Press University Press, 2005.
23. D. Ruths and L. Nakhleh. Techniques for assessing phylogenetic branch support: A performance study. In *Proceedings of the Fourth Asia-Pacific Bioinformatics Conference (APBC 06)*, pages 187–196, 2006.
24. N. Saitou and M. Nei. The neighbor-joining method: A new method for reconstructing phylogenetic trees. *Mol. Biol. Evol.*, 4:406–425, 1987.
25. M. Sanderson. `r8s` software package. Available from http://loco.ucdavis.edu/r8s/r8s.html.
26. D. L. Swofford. PAUP*: Phylogenetic analysis using parsimony (and other methods), 1996. Sinauer Associates, Underland, Massachusetts, Version 4.0.
27. F. Tajima. Evolutionary relationship of DNA sequences in finite populations. *Genetics*, 105:437–460, 1983.
28. N. Takahata. Gene genealogy in three related populations: Consistency probability between gene and population trees. *Genetics*, 122:957–966, 1989.
29. R.A. Welch, V. Burland, G. Plunkett, P. Redford, P. Roesch, D. Rasko, E.L. Buckles, S.R. Liou, A. Boutin, and J. Hackett *et al.* Extensive mosaic structure revealed by the complete genome sequence of uropathogenic Escherichia coli. *Proc. Natl. Acad. Sci. U.S.A.*, 99:17020–17024, 2002.
30. D. Zwickl and D. Hillis. Increased taxon sampling greatly reduces phylogenetic error. *Systematic Biology*, 51(4):588–598, 2002.