Integrating Sequence and Topology for Efficient and Accurate Detection of Horizontal Gene Transfer

Cuong Than, Guohua Jin, and Luay Nakhleh*

Department of Computer Science, Rice University, Houston, TX 77005, USA
nakhleh@cs.rice.edu

Abstract. One phylogeny-based approach to horizontal gene transfer (HGT) detection entails comparing the topology of a gene tree to that of the species tree, and using their differences to locate HGT events. Another approach is based on augmenting a species tree into a phylogenetic network to improve the fitness of the evolution of the gene sequence data under an optimization criterion, such as maximum parsimony (MP). One major problem with the first approach is that gene tree estimates may have wrong branches, which result in false positive estimates of HGT events, and the second approach is accurate, yet suffers from the computational complexity of searching through the space of possible phylogenetic networks.

The contributions of this paper are two-fold. First, we present a measure that computes the support of HGT events inferred from pairs of species and gene trees. The measure uses the bootstrap values of the gene tree branches. Second, we present an integrative method to speed up the approaches for augmenting species trees into phylogenetic networks.

We conducted data analysis and performance study of our methods on a data set of 20 genes from the *Amborella* mitochondrial genome, in which Jeffrey Palmer and his co-workers postulated a massive amount of horizontal gene transfer. As expected, we found that including poorly supported gene tree branches in the analysis results in a high rate of false positive gene transfer events. Further, the bootstrap-based support measure assessed, with high accuracy, the support of the inferred gene transfer events. Further, we obtained very promising results, in terms of both speed and accuracy, when applying our integrative method on these data sets (we are currently studying the performance in extensive simulations). All methods have been implemented in the PhyloNet and NEPAL tools, which are available in the form of executable code from `http://bioinfo.cs.rice.edu`.

1 Introduction

While the genetic material of an organism is mainly inherited through lineal descent from the ancestral organism, it has been shown that genomic segments in various groups of organisms may be acquired from distantly related organisms through *horizontal DNA, or gene, transfer* (HGT). It is believed that HGT is ubiquitous among prokaryotic organisms (24; 6) and plays a significant role in their genomic diversification (20). Recent studies have also demonstrated evidence of massive HGT in various groups of plants (3; 4).

* Corresponding author.

C.E. Nelson and S. Vialette (Eds.): RECOMB-CG 2008, LNBI 5267, pp. 113–127, 2008.

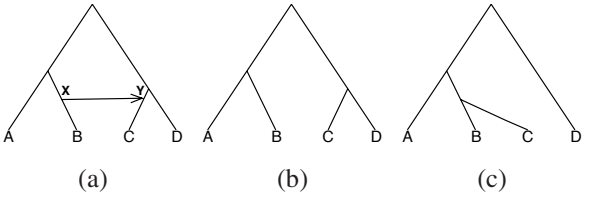


Fig. 1. (a) A phylogenetic network with a single HGT edge from X to Y, ancestors of taxa B and C, respectively. (b) The underlying species tree, which models the evolution of vertically transmitted genomic regions. (c) The tree that models the evolution of horizontally transferred genomic regions. The phylogenetic network *contains* the two trees, in the sense that each of the trees can be obtained from the network by removing all but one incoming edge (branch) for each node in the network.

When HGT occurs among organisms, the evolution of their genomes is best modeled as a *phylogenetic network* (19; 15) which can be viewed as the reconciliation of the evolutionary trees of the various genomic segments, typically referred to as *gene trees*. Figure 1(a) shows an example of a phylogenetic network on four taxa in the presence of an HGT from X, an ancestor of taxon B, to Y, an ancestor of taxon C. In this scenario, genomic regions that are not involved in the horizontal transfer are inherited from the ancestral genome, and their evolution is modeled by the tree in Figure 1(b), while horizontally transferred regions in C's genome are acquired from X, and their evolution is modeled by the tree in Figure 1(c).

As illustrated in Figure 1, the occurrence of HGT may result in gene trees whose topologies are discordant (the trees in (b) and (c)). Detecting discordance among gene trees, particularly with respect to a species tree, and reconciling them into a phylogenetic network, are the fundamental principles on which phylogeny-based HGT detection approaches are built. Several algorithms and software tools based on this approach have been introduced recently (e.g., (8; 22; 2)), all of which infer HGT events by comparing the topologies (shapes) of trees.

A major confounding factor that negatively affects the accuracy of phylogeny-based methods is that reconstructed species/gene trees usually contain error in the form of wrong branches. These errors result in species/gene tree incongruities, thus triggering phylogeny-based methods to make false predictions of HGT events (29). For example, the incongruence between the two tress in Figure 1 could have arisen simply due to poor reconstruction of the gene tree, even though no HGT was involved. In general, the performance of phylogeny-based methods, in terms of both accuracy and speed, is negatively affected by several factors, which include errors in the estimated trees and the exponential space of potential solutions (28; 29). In this paper we address this confounding factor by devising a method that estimates the support of inferred HGT events by using the bootstrap values of the gene tree branches.

Recently, a set of methods were devised to estimate HGT events by augmenting a species tree into a phylogenetic network to improve the fitness of evolution of the gene's sequence data based on the maximum parsimony (MP) criterion (21; 13). While yielding promising results, these methods were very slow, since the problem of inferring a phylogenetic network under the MP criterion is NP-hard and even hard to approximate

(11; 14). In this paper, we present a new integrative method for improving the speed of the MP-based methods by first conducting a topology-based analysis, based on the topology of the tree inferred from the gene's sequences, and then screening the inferred events based on the MP criterion.

We have implemented both approaches and analyzed a data set of 20 genes that exhibited massive HGT in the basal angiosperm *Amborella* according to (4). First, we demonstrated the effects of error in the reconstructed gene trees on the estimates of gene transfer. We found that including poorly supported gene tree branches in the analysis results in a high rate of false positives. Second, the support measure assessed, with high accuracy, the support of the gene transfers inferred by the topology-based analysis. Third, the integrative method, that combines topology comparison with the MP criterion, detected efficiently all but one of HGT edges that were postulated by the authors. Further, our approach detected new candidate HGTs, with high support. The combined accuracy and efficiency of our approach, achieved by integrating topological analysis with sequence information, will enable automated detection of HGT in large data sets.

2 Materials and Methods

2.1 Topology-Based HGT Detection and the RIATA-HGT Tool

When HGT occurs, the evolutionary history of the genomes may be more appropriately represented by a *phylogenetic network*, which is a rooted, directed, acyclic graph (rDAG) that extends phylogenetic trees to allow for nodes with more than a single parent (1; 19; 15; 9). The phylogeny-based HGT reconstruction problem seeks the phylogenetic network with minimum number of HGT edges (equivalently, the minimum number of nodes that have more than a single parent), to reconcile the species and gene trees. The minimization requirement simply reflects a parsimony criterion: in the absence of any additional biological knowledge, the simplest solution is sought. In this case, the simplest solution is one that invokes the minimum number of HGT events to explain tree incongruence.

Definition 1. *(The HGT Detection Problem)*

> **Input:** *Species tree ST and gene tree GT, both of which are rooted.*
> **Output:** *A set with the fewest directed HGT edges, each of which is posited between a pair of branches of ST, such that the resulting phylogenetic network contains the gene tree GT.*

As indicated above, the resulting network is an rDAG, where the network's branches that are also branches in ST are directed away from the root, and the HGT edges are directed as indicated by their description. For example, the phylogenetic network in Figure 1(a) is a solution to the HGT Detection Problem if we consider the trees in 1(b) and 1(c) to be the species and gene trees, respectively, since it has a smallest set of HGT edges (one, in this case) and contains the gene tree. Notice that in this case the gene tree cannot be reconciled with the species tree without HGT edges.

Several heuristics for solving the problem have been recently introduced, e.g., (18; 8; 7; 23; 17; 2; 10; 22). As mentioned above, the performance of methods following this

approach, in terms of both accuracy and speed, is negatively affected by several factors, which include errors in the estimated trees and the exponential space of potential solutions (28; 29). We have recently addressed these issues and extended the RIATA-HGT method (22) so that it detects HGT in pairs of trees that are not necessarily binary, and computes multiple minimal solutions very efficiently (27).

2.2 Assessing the Support of HGT Edges

As mentioned above, topology-based HGT detection methods are sensitive to error in the inferred trees, as illustrated on simulated data sets and reported in (28; 29) and on biological data sets, which we report here. In this paper, we propose a measure of the *support* of an inferred HGT edge based on the bootstrap values of the gene tree branches. Roughly speaking, the support value of HGT edge $X \rightarrow Y$ in the species tree, where Y' is the sibling of Y, is derived from the bootstrap values of the gene tree branches that separate the clade under Y from the clade under Y'. The rationale behind the idea is that if Y' and Y are well separated in the gene trees (i.e., some branches in the path from Y to Y' have high bootstrap values), an HGT is necessary to move Y away from Y'). For example, the support of HGT edge $X \rightarrow Y$ in Figure 2 is calculated based on the bootstrap values of the branches separating B from A in the gene tree.

However, since trees are not necessarily binary in general, and given that multiple HGT edges may involve branches under X or Y, the calculation is more involved (e.g., see Figure 3) and requires a formal treatment, which we now provide.

Given a species tree ST, a gene tree GT, and a set Ξ of HGT edges, create a network N by adding every edge in Ξ to ST. We create two trees ST' and ST'' from N, as follows:

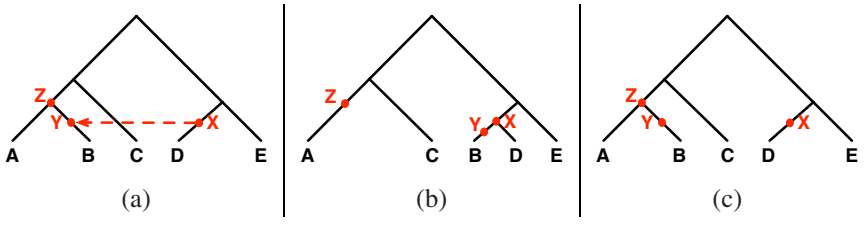


Fig. 2. An illustration of computing the support value of an HGT edge. In this case, the support of HGT edge $X \rightarrow Y$ added on the species tree, resulting in the network in (a), is calculated based on the bootstrap of the branches that separate the "moving clade" rooted at Y (which, in this case, is the clade that contains the single leaf B) from its sister clade (which, in this case, is the clade that contains the single leaf A) in the gene tree (b). The species tree is depicted by the solid lines in (a), and the gene tree by the solid lines in (b). The tree in (b), along with all internal nodes, including nodes Y and Z, is the tree ST' used in the procedure for computing the support value, whereas the tree in (c), along with all internal nodes, including nodes X and Y, is the tree ST'' used in the procedure. Notice that nodes Y and Z in (b), as well as nodes X and Y in (c) have in-degree and out-degree 1.

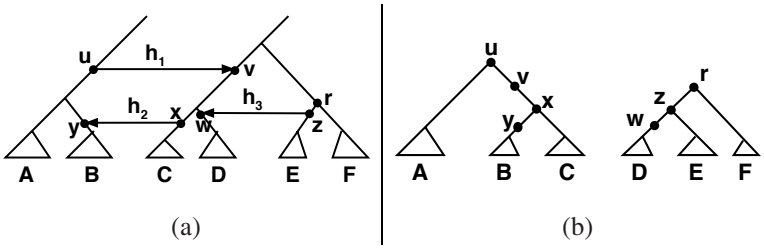


Fig. 3. An illustration of calculating support values of HGT edges when multiple HGT edges are involved. The HGT edges (arrows) h_1, h_2, and h_3, are posited between pairs of branches (lines) in the species tree ST resulting in the phylogenetic network N, part of which is shown in (a). When a genomic segment is transferred across all these three HGT edges, the evolutionary history of that segment in the six taxa A, B, C, D, E, and F is represented by the two clades in (b). Further, these two clades and their internal nodes are part of the tree ST' that is generated in the first step of the support calculation. The tree ST'' is obtained from the tree ST in (a) by adding the endpoints of all HGT edges (the solid circles), but removing the HGT edges themselves. The support of h_1 is derived from the bootstrap support of the gene tree branches that separate the MRCA of set P of leaves (which, in this case, is the set of leaves in clades B and C combined) and the MRCA of set Q of leaves (which, in this case, is the set of leaves in clades D, E, and F combined). The support of h_2 is derived from the bootstrap values of the gene tree branches that separate the MRCA of set P of leaves (in this case, the leaves of clade B) and the MRCA of set Q of leaves (in this case, the leaves of clade A). The support of h_3 is derived from the bootstrap support of the gene tree branches that separate the MRCA of set P of leaves (in this case, the leaves of clade D) and the MRCA of set Q of leaves (in this case, the leaves in clades B and C combined).

- A tree ST' is built from N in such a way that for each edge $X \rightarrow Y$ in Ξ, this edge is kept, and the other edge incident into Y is deleted. The tree in Figure 2(b), including nodes Y and Z is the tree ST' obtained from the network in Figure 2(a).
- A tree ST'' is built from N in a similar fashion, but edge $X \rightarrow Y$ is deleted while the other edge incident into Y is kept. The tree in Figure 2(c), including nodes X and Y is the tree ST'' obtained from the network in Figure 2(a).

Notice that both trees ST' and ST'' have nodes of in-degree and out-degree 1. Retaining these nodes in these two trees ensures the well-definedness of the procedure that we will describe below for computing the support of HGT edge $X \rightarrow Y$. Note that ST' and ST'' can have nodes whose in-degree and out-degree are both 1. One important fact about ST' and ST'' that is necessary for our method for assessing HGT support is that they have the same set of nodes. See Figure 3 for an illustration. We denote by $L_T(v)$ the set of leaves under node v in a tree T; i.e., the set of leaves to which the paths from the root of T must go through node v. We define the support of an HGT edge $h = X \rightarrow Y$, which we denote by $b(h)$, assuming the bootstrap support of the gene tree branches have been computed.

Our task is to find the path of edges between the "moving clade" (the clade of taxa this is "moved" in the HGT event) and its sister clade in the gene tree. In Figure 2, $P = \{B\}$ and $Q = \{A\}$. We define two sets of leaves: P, which is the leaf-set of

the moving clade, and Q, which is the leaf-set of the moving clade's sister clade. More formally, $P = L_{ST'}(Y)$, and Q is defined as follows:

1. Let $Y' = Y$.
2. Let node p be the parent of Y' in ST''.
   (a) If $L_{ST'}(p) \neq \emptyset$, then $Q = L_{ST'}(p)$.
   (b) Else, let $Y' = p$, and go back to step (2).

To illustrate these sets from Figure 3:

- For HGT edge h_1: P contains only the leaves in clades B and C, and Q contains only the leaves in clades D, E and F.
- For HGT edge h_2: P contains only the leaves in clade B, and Q contains only the leaves in clade A.
- For HGT edge h_3: P contains only the leaves in clade D, and Q contains only the leaves in clades B and C.

In this example, we get the set Q after only one iteration of the procedure. The need for repeating Step 2 lies in the case where all siblings of Y are moved by HGT events.

Now that we have computed the sets P and Q of leaves, let p and q be the most recent common ancestor nodes of P and Q, respectively, in the gene tree. Let $\mathcal{E}$ be the set of branches in the gene tree between nodes p and q. The support value of the HGT edge h is

$$b(h) = \max_{e \in \mathcal{E}} s(e), \tag{1}$$

where $s(e)$ is the bootstrap value of the branch e in the gene tree. We choose to use the maximum bootstrap value of a branch on the path since that value alone determines whether the donor and recipient involved in an HGT edge truly form a clade in the gene tree. Notice that averaging all values may not work, since, for example, if the path has many branches, only one of which has very high support and the rest have very low support, averaging would reflect low support for the HGT edge.

In case the species tree branches have bootstrap support values associated with them, these can also be incorporated as follows for HGT edge $h : X \rightarrow Y$. Let Z be the MRCA of X and Y in the species tree, and let $\mathcal{E}'$ be the set of branches in the species tree on the paths from Z to X and from Z to Y. Then, Formula (1) can be modified to become

$$b(h) = \min\{\max_{e \in \mathcal{E}} s(e), \max_{e' \in \mathcal{E}'} s(e')\}, \tag{2}$$

where $s(e)$ is the bootstrap value of the gene tree branch for $e \in \mathcal{E}$ and the species tree branch for $e' \in \mathcal{E}'$.

2.3 Parsimony-Based HGT Detection and the NEPAL Tool

The relationship between a phylogenetic network and its constituent trees is the basis for the MP extension to phylogenetic networks described in a sequence of papers by Jin, Nakhleh and co-workers (21; 11; 13; 14), which we now review briefly.

The Hamming distance between two equal-length sequences x and y, denoted by $H(x, y)$, is the number of positions j such that $x_j \neq y_j$. Given a fully-labeled tree

T, i.e., a tree in which each node v is labeled by a sequence s_v over some alphabet Σ, we define the Hamming distance of a branch $e \in E(T)$ ($E(T)$ denotes the set of all branches in tree T), denoted by $H(e)$, to be $H(s_u, s_v)$, where u and v are the two endpoints of e. We now define the parsimony length of a tree T.

Definition 2. *The parsimony length of a fully-labeled tree T, is $\sum_{e \in E(T)} H(e)$. Given a set S of sequences, a maximum parsimony tree for S is a tree leaf-labeled by S and assigned labels for the internal nodes, of minimum parsimony length.*

Given a set S of sequences, the MP problem is to find a maximum parsimony phylogenetic tree T for the set S. The evolutionary history of a single (non-recombining) gene is modeled by one of the trees contained inside the phylogenetic network of the species containing that gene. Therefore the evolutionary history of a site s is also modeled by a tree contained inside the phylogenetic network. A natural way to extend the tree-based parsimony length to fit a dataset that evolved on a network is to define the parsimony length for each site as the minimum parsimony length of that site over all trees contained inside the network.

Definition 3. *The parsimony length of a network N leaf-labeled by a set S of sequences, is*

$$NCost(N, S) := \sum_{s_i \in S} (\min_{T \in \mathcal{T}(N)} TCost(T, s_i))$$

where $TCost(T, s_i)$ is the parsimony length of site s_i on tree T.

Notice that as usually large segments of DNA, rather than single sites, evolve together, Definition 3 can be extended easily to reflect this fact, by partitioning the sequences S into non-overlapping blocks b_i of sites, rather than sites s_i, and replacing s_i by b_i in Definition 3. This extension may be very significant if, for example, the evolutionary history of a gene includes some recombination events, and hence that evolutionary history is not a single tree. In this case, the recombination breakpoint can be detected by experimenting with different block sizes. Based on this criterion, we would want to reconstruct a phylogenetic network whose parsimony length is minimized. In the case of horizontal gene transfer, a species tree that models vertical inheritance is usually known; e.g., see (16). Hence, the problem of reconstructing phylogenetic networks in this case becomes one of finding a set of edges whose addition to the species tree "best explains" the horizontal gene transfer events, which is defined as the θ-FTMPPN problem in (13). We have implemented heuristics to solve this problem in the NEPAL software tool.

Jin *et al.* demonstrated the high accuracy of HGT detection by solving the θ-FTMPPN problem on a wide array of simulated and biological data sets (13). However, the major drawback of the approach was the high running time (several hours to days for detecting a small number of HGT events on small species trees). In the next section, we propose an approach that integrates topological comparison of trees with the MP criterion to achieve both accuracy and computational efficiency in the detection of HGT events.

2.4 Integrating MP and Topological Comparison

As described above, the extended version of RIATA-HGT performs very well, in terms of speed, but overestimates the number of HGT events (due to inaccuracies in the trees).

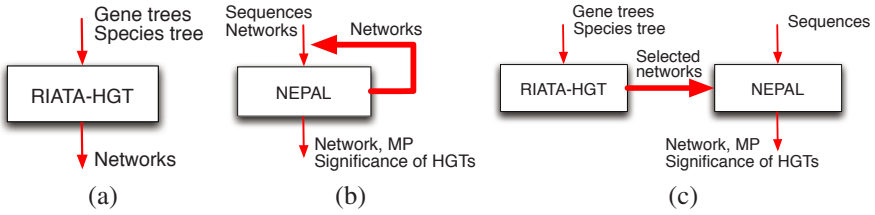


Fig. 4. Diagrams of the three approaches: RIATA-HGT (a), NEPAL (b), and the integrative approach (c). RIATA-HGT takes as input a pair of trees (species and gene trees), and computes a set of minimal phylogenetic networks that contain both trees. NEPAL takes as input a set of sequences and a set of networks (initially, the set of networks includes only the species tree), and solves the θ-FTMPPN problem by finding a set of HGT edges whose addition to the species tree results in an optimal phylogenetic network under the maximum parsimony criterion. In our analysis in this paper, the value of θ was determined based on observing the parsimony improvement, and is data set-dependent. The integrative approach first runs RIATA-HGT on the pair of species/gene trees, and then uses the phylogenetic networks inferred to guide the search of the parsimony search, as implemented in NEPAL.

On the other hand, Jin *et al.* have shown that the MP approach performs very well in terms of accuracy, yet is very slow (13). The computational requirement of the MP criterion stems from the large space of networks that the method has to consider. To achieve good accuracy with high efficiency, we propose to integrate the two approaches.

Figure 4 shows the diagrams of the two separate approaches as well as the proposed integrative approach. In contrast to the MP approach, where a large number of networks is explored due to the exponential number of combinations of candidate HGT edges, our integrative approach focuses on the phylogenetic networks inferred by RIATA-HGT. The rationale behind this approach is that the overestimation of RIATA-HGT, and topology-based HGT detection methods in general, is mainly in the form of *false positives*, whereas the *false negatives* are negligible. In other words, RIATA-HGT in most cases infers all the "true" HGT edges, but also infers additional "false" HGT edges due to errors in the gene trees. Given the accuracy of the MP criterion, we post-process the results of RIATA-HGT by evaluating them based on the MP criterion, with the hope that it would remove the false positive HGT edges. We have developed efficient algorithms for evaluating the parsimony length of phylogenetic networks (14), so both steps of our integrative approach can be carried out efficiently. The algorithm of the integrative approach is outlined in Figure 5.

The time complexity of our integrative algorithm is $O(km2^k)$ using exact network parsimony length computing algorithms and $O(mk^2)$ using the approximation algorithm described in (14).

2.5 Validating the Integrative Approach

To validate the integrative approach, we consider the differences in the HGT edges and parsimony scores computed by each of the two methods. More formally, if N_{int} and N_{mp} are two phylogenetic networks computed by the integrative and MP approaches, respectively, on the same species trees ST and set of sequences S, and $H(N_{\text{int}})$ and

TopSeqHGTIdent(ST, GT, S)
INPUT: species tree ST, gene tree GT, and sequence dataset S
OUTPUT: network N with marked significance of each HGT

1 Let $\{N_1, \ldots, N_m\}$ be the set of all phylogenetic networks computed by RIATA-HGT, and let $H(N_i)$ be the set of HGT edges in N_i.
2 Let $\mathcal{H} = \cap_{i=1}^{m} H(N_i)$, and $R(N_i) = H(N_i) - \mathcal{H}$. In other words, $\mathcal{H}$ denotes the set of HGT edges that are shared by all networks, and $R(N_i)$, for $1 \leq i \leq m$, the set of HGT edges that are in N_i but not shared by all other networks.
3 Apply NEPAL to $N' = ST + \mathcal{H}$.
4 For each network N_i, $1 \leq i \leq m$, apply NEPAL by incrementally adding (in no particular order) the HGT edges in $R(N_i)$ to N', and compute the minimum parsimony length of the phylogenetic network.
5 Let $N = ST$, N_{opt} be the best network according to maximum parsimony criterion, that is $NCost(N_{opt}, S) = min_{i=1}^{m}(NCost(N_i, S))$. Apply NEPAL by adding to ST each time one of the HGT events $h \in H(N_{opt})$ that results in the most significant drop in the parsimony score and let $N = N \cup h$. Stop this process when the drop is smaller than a specified threshold.

Fig. 5. An outline of the integrative approach. Steps 1 through 4 seek the most parsimonious network among the several networks computed by RIATA-HGT. Step 5 conducts one last pass of parsimony length calculation to identify the HGT edges in the optimal network whose contribution to lowering the parsimony length is significant. This last step is necessary since, even though the most parsimonious network is optimal among all networks computed by RIATA-HGT, it may still have " parsimoniously unnecessary" HGT edges whose addition by RIATA-HGT was necessitated by the incongruence between the two trees.

$H(N_{\rm mp})$ are the sets of HGT edges in $N_{\rm int}$ and $N_{\rm mp}$ respectively, then we have two measures of quality:

- $m^{\rm HGT}(N_{\rm int}, N_{\rm mp}) = (H(N_{\rm int}) - H(N_{\rm mp})) \cup (H(N_{\rm mp}) - H(N_{\rm int}))$. This measure reflects the difference in the locations of the HGT edges between the two networks.
- $m^{\rm pars}(N_{\rm int}, N_{\rm mp}) = |NCost(N_{\rm int}, S) - NCost(N_{\rm mp}, S)|$. This measure quantifies the difference in the MP scores of the two networks.

2.6 Data

We studied 20 out of the 31 mitochondrial gene data sets, which were collected from 29 diverse land plants and analyzed in (4). These are *cox2*, *nad2*, *nad3*, *nad4(ex4)*, *nad4(exons)*, *nad5*, *nad6*, *nad7*, *atp1*, *atp8*, *ccmB*, *ccmC*, *ccmFN1*, *cox3*, *nad1*, *rpl16*, *rps19*, *sdh4*, and three introns *nad2intron*, *nad5intron* and *nad7intron*. We used a species tree for the dataset based on information at NCBI (`http://www.ncbi.nih.gov`) and analyzed the entire dataset with both seed and nonseed plants together. For each gene data set, we restricted the species tree to those species for which the gene sequence is available. For the parsimony analysis, we analyzed each gene data set separately, by solving the θ-FTMPPN problem, as implemented in NEPAL, on the gene DNA sequence with respect to the species tree. The θ-FTMPPN problem seeks a set of HGT edges, each of whose addition to the species tree improves the parsimony score

beyond the specified threshold θ. In this paper, we determined the value of θ based on observing the parsimony improvement, as HGT edges were added, and the value of θ was dependent on the data set; i.e., data sets did not necessarily have the same value of θ. Our current criterion for determining the value of θ is very simple, yet shows very good accuracy: we observe the slope of the decrease in the parsimony score as the HGT edges are added, and stop adding new edges when the slope changes (slows down) significantly. In all data sets we have analyzed so far, the change in slope has been very sharp, which makes it very straightforward to determine the number of HGT edges to add. For the analysis by RIATA-HGT (which takes as input a pair of trees), for each gene, we used the species tree as the reference tree and the gene tree reported in (4) as the second tree. Bergthorsson *et al.* also calculated and reported the bootstrap support of the gene tree branches, which we used in calculating the support values of HGT edges inferred by RIATA-HGT. For the integrative approach, we used the phylogenetic networks produced by RIATA-HGT and the gene sequences, and applied the algorithm described above. It is important to note that in their analyses, Bergthorsson *et al.* focused only on genes that were horizontally transferred to the mitochondrial genome of *Amborella*.

3 Results and Discussion

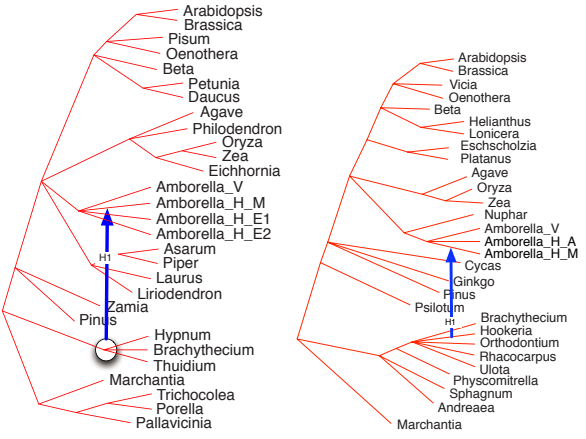
Table 1 summarizes information about the HGT edges reported by Bergthorsson *et al.* for the 20 genes (4), the HGT edges inferred by the MP criterion as implemented in the NEPAL tool (13; 14), the HGT edges inferred by the RIATA-HGT method for topology-based HGT detection (22; 27), and the HGT edges inferred by the integrative approach of topology- and MP-based analysis. Figure 6 shows the HGT edges inferred for two data sets; we omit the phylogenetic networks for the other data sets.

Bergthorsson *et al.* reported the groups of species to which the donor(s) of horizontally transferred genes belong, rather than the specific donor. In particular, they focused on four groups: Bryophytes, Moss, Eudicots, and Angiosperms. For the recipient, the authors only focused on *Amborella*. Further, for each HGT event (edge), they computed and reported its significance based on the SH test (25). Of the 25 HGT events that Bergthorsson *et al.* postulated, 13 were supported, 9 unsupported, and 3 (the 3 intron data sets) had no reported support.

Of the 13 HGTs reported in (4) with high support according to the SH test, the MP analysis identified 12, missing only the HGT involving gene *nad5* from the Angiosperms. RIATA-HGT also identified 12 out of the 13 HGT edges, missing only the HGT involving gene *nad6*. Further, all these 12 HGT edges identified by RIATA-HGT had support (out of 100) higher than 95, based on Formula (1) above. The integrative approach identified 11 of the 13 well-supported HGT edges, missing only the HGTs involving genes *nad5* and *nad6*. In the former case, the HGT was identified by RIATA-HGT, but deemed insignificant in the parsimony analysis phase, and in the latter case, RIATA-HGT missed the HGT edge, and since the parsimony analysis in the integrative approach uses those edges that RIATA-HGT identifies, this HGT edge was not identified.

Table 1. Mitochondrial gene data sets and HGTs postulated in (4) and those computed by the MP analysis (NEPAL), RIATA-HGT, and the integrative approach (RIATA-HGT+MP). The '#HGTs' value is the number of HGTs found; 'donor' denotes the group from which the gene was transferred (in all cases, the recipient is *Amborella*; 'SH' denotes support of the HGT events as computed by the SH test (25) and reported in (4) (values lower than 0.05 indicate high support, and NS indicates support is not significant). The 'F?' column indicates which of the HGTs postulated by the authors were found by the MP analysis, and the '#Nets' shows the number of networks analyzed by the parsimony method to identify the correct HGTs. For the RIATA-HGT column, each row shows the minimum number of HGTs computed, the number of minimal solutions, and the number of distinct HGTs in all minimal solutions, respectively. B=Bryophyte, M=Moss, E=Eudicot, and A=Angiosperm.

	Bergthorsson *et al.*			MP			RIATA-HGT			RIATA-HGT+MP		
Gene	#HGTs	donor	SH	#HGTs	F?	#Nets	#HGTs	#Nets	#events	#HGTs	F?	#Nets
cox2	3	M	<0.001	1	Y	8482	9	4	12	1	Y	23
		E	NS		N	8482	9	4	12		N	23
		E	NS		N	8482	9	4	12		N	23
nad2	2	M	<0.001	1	Y	3500	7	6	11	1	Y	21
		E	NS		N	3500	7	6	11		N	21
nad4(exons)	1	M	<0.001	1	Y	1620	4	2	5	1	Y	9
nad4(ex4)	1	E	NS	2	Y	1832	6	3	8	2	Y	21
nad5	2	M	<0.001	1	Y	3292	6	6	9	1	Y	17
		A	0.025		N	3292	6	6	9		N	17
nad6	1	B	<0.001	1	Y	2484	6	3	8	1	N	15
nad7	2	M	<0.001	1	Y	2948	7	1	7	1	Y	13
		E	NS		N	2948	7	1	7		N	13
atp1	1	E	0.001	1	Y	2817	6	18	14	1	Y	27
atp8	1	E	0.008	2	Y	9059	5	6	11	1	Y	21
ccmB	1	E	NS	2	Y	66015	6	3	14	2	Y	101
ccmC	1	E	0.03	1	Y	2786	7	21	15	1	Y	29
ccmFN1	1	E	0.004	2	Y	4412	7	18	13	2	Y	46
cox3	1	A	NS	1	N	3466	8	15	18	1	N	52
nad1	1	E	<0.001	1	Y	2812	9	12	14	1	Y	27
rpl16	1	E	NS	3	Y	21632	10	27	23	1	Y	67
rps19	1	E	0.003	1	Y	1476	5	4	7	1	Y	13
sdh4	1	E	NS	3	Y	18670	9	18	18	3	Y	540
nad2intron	1	M	–	2	Y	5904	8	2	10	2	Y	44
nad5intron	1	M	–	2	Y	10280	9	5	18	2	Y	51
nad7intron	1	M	–	1	Y	3284	12	48	26	1	Y	51



The *cox2* gene data set The *nad5* gene data set

Fig. 6. HGT edges identified computationally. In the *cox2* data set, the MP analysis identified four equally "good" HGT edges, all of which represent HGT to Amborella_H_M, and each of which denotes a different donor (the four branches inside the circle): Hypnum, Brachythecium, Thuidium, and the clade containing all three. Under the MP criterion, those four donors contribute equally to the improvement in the parsimony length. RIATA-HGT identified the HGT edge with the clade of all three species as the donor, and Amborella_H_M as the recipient, along with many other "false positive" HGT edges, which are not shown here. The integrative approach finds the single HGT edge from the clade of all three species to Amborella_H_M. In the case of the *nad5* data set, the MP analysis identified the only shown HGT edge, which was identified as well by RIATA-HGT along with other HGT edges (not shown). The integrative approach identified only the shown HGT edge.

The only HGT that the MP analysis missed in the *nad5* data set involves an HGT from the Angiosperms. This transfer had support of 0.025 based on the SH test as reported by Bergthorsson *et al.*. Further, this edge was identified by RIATA-HGT with support of 94. In the MP analysis, the only significant *Amborella* transfer comes from the Moss group. The grouping of Eudicots and Monocots results in much less significant improvement in the parsimony length. The transfer from the Angiosperms to *Amborella* has even less impact on the parsimony length.

The three HGT edges postulated by Bergthorsson *et al.* for the intron data sets, and which had no support values based on the SH test reported, were identified by the MP analysis of NEPAL, RIATA-HGT, and the integrative approach. Further, those three edges had support around 50, based on Formula (1) above.

Of the other 9 HGT events reported by the authors with no significant support based on the SH test, the MP analysis, RIATA-HGT and the integrative approach did not identify five of them. However, all three approaches identified four HGT edges that had no significant support based on the SH test. All these four HGTs were from the Eudictots to *Amborella*, and they were in the *nad4(ex4)*, *ccmB*, *rpl16*, and *sdh4* data sets. In all four cases, the support ranged from low ($\approx$ 50) to high ($\approx$ 95), based on Formula (1) above.

In eight data sets, the MP analysis identified HGT edges in addition to those reported in (4). However, none of these edges involved Amborella. One possible explanation for why these edges were not reported in (4) is probably because the authors focused only on HGT events involving *Amborella*. Another explanation may be the inaccuracy of the parsimony criterion for evaluating HGT edges in these cases. RIATA-HGT identified the same HGT edges in these eight cases, support ranging from low (≈ 50) to high (≈ 90), based on Formula (1) above. The integrative approach in these cases also identified the same HGT edges.

It is important to note that all other HGT edges identified by RIATA-HGT had support ranging from very low to very high (≈ 100 in some cases), based on Formula (1) above, but were rendered insignificant based on the parsimony phase in the integrative approach as well. Further, we analyzed the parsimony scores of the networks computed by NEPAL and by the integrative approach. In both cases, the scores were very similar.

Finally, the parsimony approach analyzed several thousand networks (using the best available heuristic (11)) to identify the HGTs, and that took several hours on each gene, and up to two days on some. Table 1 shows the exact number of networks that our parsimony analysis checks even with the branch-and-bound heuristics. The integrative approach, on the other hand, took a few minutes on each of the data sets.

4 Conclusions and Future Work

In this paper, we presented a measure to assess the support of HGT edges inferred by phylogeny-based methods that compare species and gene tree topologies. Further, we presented a method for speeding up approaches the infer HGT events by augmenting species trees into phylogenetic networks to improves the fitness of evolution of gene sequence data. We obtained promising results on 20 mitochondrial gene data sets, previously analyzed in (4).

While we used the maximum parsimony criterion in the second phase of our integrative approach, it is also possible to use stochastic models of HGT (e.g., (26; 12; 5)) to probabilistically screen the phylogenetic networks produced in the first phase. This is one of the future directions we intend to pursue. The immediate task for us in our future work is to study the performance of these measures in extensive simulations. The advantage that simulations provide over real data is that the true HGT events are known, which allows us to make absolute quantification of the performance of the methods.

Acknowledgments

This work is supported in part by the Department of Energy grant DE-FG02-06ER25734, the National Science Foundation grant CCF-0622037, and the George R. Brown School of Engineering Roy E. Campbell Faculty Development Award. Further, the work was supported in part by the Rice Computational Research Cluster funded by NSF under Grant CNS-0421109, and a partnership between Rice University, AMD and Cray.

The authors would like to thank Aaron O. Richardson for providing the *Amborella* gene data sets, and the anonymous reviewers for comments on the technical details as well as the readability of the manuscript.

References

[1] Baroni, M., Semple, C., Steel, M.: A framework for representing reticulate evolution. Annals of Combinatorics 8(4), 391–408 (2004)
[2] Beiko, R.G., Hamilton, N.: Phylogenetic identification of lateral genetic transfer events. BMC Evolutionary Biology 6 (2006)
[3] Bergthorsson, U., Adams, K.L., Thomason, B., Palmer, J.D.: Widespread horizontal transfer of mitochondrial genes in flowering plants. Nature 424, 197–201 (2003)
[4] Bergthorsson, U., Richardson, A., Young, G.J., Goertzen, L., Palmer, J.D.: Massive horizontal transfer of mitochondrial genes from diverse land plant donors to basal angiosperm Amborella. Proc. Nat'l Acad. Sci., USA 101, 17747–17752 (2004)
[5] Galtier, N.: A model of horizontal gene transfer and the bacterial phylogeny problem. Systematic Biology 56(4), 633–642 (2007)
[6] Gogarten, J.P., Doolittle, W.F., Lawrence, J.G.: Prokaryotic evolution in light of gene transfer. Mol. Biol. Evol. 19(12), 2226–2238 (2002)
[7] Gorecki, P.: Reconciliation problems for duplication, loss and horizontal gene transfer. In: Proc. 8th Ann. Int'l Conf. Comput. Mol. Biol. (RECOMB 2004), pp. 316–325 (2004)
[8] Hallett, M.T., Lagergren, J.: Efficient algorithms for lateral gene transfer problems. In: Proc. 5th Ann. Int'l Conf. Comput. Mol. Biol. (RECOMB 2001), pp. 149–156. ACM Press, New York (2001)
[9] Huson, D.H., Bryant, D.: Application of phylogenetic networks in evolutionary studies. Molecular Biology and Evolution 23(2), 254–267 (2006)
[10] Huson, D.H., Kloepper, T., Lockhart, P.J., Steel, M.A.: Reconstruction of reticulate networks from gene trees. In: Miyano, S., Mesirov, J., Kasif, S., Istrail, S., Pevzner, P.A., Waterman, M. (eds.) RECOMB 2005. LNCS (LNBI), vol. 3500, pp. 233–249. Springer, Heidelberg (2005)
[11] Jin, G., Nakhleh, L., Snir, S., Tuller, T.: Efficient parsimony-based methods for phylogenetic network reconstruction. Bioinformatics 23, e123–e128 (2006); Proceedings of the European Conference on Computational Biology (ECCB 2006)
[12] Jin, G., Nakhleh, L., Snir, S., Tuller, T.: Maximum likelihood of phylogenetic networks. Bioinformatics 22(21), 2604–2611 (2006)
[13] Jin, G., Nakhleh, L., Snir, S., Tuller, T.: Inferring phylogenetic networks by the maximum parsimony criterion: a case study. Molecular Biology and Evolution 24(1), 324–337 (2007)
[14] Jin, G., Nakhleh, L., Snir, S., Tuller, T.: A new linear-time heuristic algorithm for computing the the parsimony score of phylogenetic networks: theoretical bounds and empirical performance. In: Măndoiu, I.I., Zelikovsky, A. (eds.) ISBRA 2007. LNCS (LNBI), vol. 4463, pp. 61–72. Springer, Heidelberg (2007)
[15] Kunin, V., Goldovsky, L., Darzentas, N., Ouzounis, C.A.: The net of life: reconstructing the microbial phylogenetic network. Genome Research 15, 954–959 (2005)
[16] Lerat, E., Daubin, V., Moran, N.A.: From gene trees to organismal phylogeny in prokaryotes: The case of the γ-proteobacteria. PLoS Biology 1(1), 1–9 (2003)
[17] MacLeod, D., Charlebois, R.L., Doolittle, F., Bapteste, E.: Deduction of probable events of lateral gene transfer through comparison of phylogenetic trees by recursive consolidation and rearrangement. BMC Evolutionary Biology 5 (2005)
[18] Makarenkov, V.: T-REX: Reconstructing and visualizing phylogenetic trees and reticulation networks. econstructing and visualizing phylogenetic trees and reticulation networks 17(7), 664–668 (2001)
[19] Moret, B.M.E., Nakhleh, L., Warnow, T., Linder, C.R., Tholse, A., Padolina, A., Sun, J., Timme, R.: Phylogenetic networks: modeling, reconstructibility, and accuracy. IEEE/ACM Transactions on Computational Biology and Bioinformatics 1(1), 13–23 (2004)

[20] Nakamura, Y., Itoh, T., Matsuda, H., Gojobori, T.: Biased biological functions of horizontally transferred genes in prokaryotic genomes. Nature Genetics 36(7), 760–766 (2004)
[21] Nakhleh, L., Jin, G., Zhao, F., Mellor-Crummey, J.: Reconstructing phylogenetic networks using maximum parsimony. In: Proceedings of the 2005 IEEE Computational Systems Bioinformatics Conference (CSB 2005), pp. 93–102 (2005)
[22] Nakhleh, L., Ruths, D., Wang, L.S.: RIATA-HGT: A fast and accurate heuristic for reconstrucing horizontal gene transfer. In: Wang, L. (ed.) COCOON 2005. LNCS, vol. 3595, pp. 84–93. Springer, Heidelberg (2005)
[23] Nakhleh, L., Warnow, T., Linder, C.R.: Reconstructing reticulate evolution in species–theory and practice. In: Proc. 8th Ann. Int'l Conf. Comput. Mol. Biol. (RECOMB 2004), pp. 337–346 (2004)
[24] Ochman, H., Lawrence, J.G., Groisman, E.A.: Lateral gene transfer and the nature of bacterial innovation. Nature 405(6784), 299–304 (2000)
[25] Shimodaira, H., Hasegawa, M.: Multiple comparisons of log-likelihoods with applications to phylogenetic inference. Molecular Biology and Evolution 16, 1114–1116 (1999)
[26] Suchard, M.A.: Stochastic models for horizontal gene transfer: taking a random walk through tree space. Genetics 170, 419–431 (2005)
[27] Than, C., Nakhleh, L.: SPR-based tree reconciliation: Non-binary trees and multiple solutions. In: Proceedings of the Sixth Asia Pacific Bioinformatics Conference, pp. 251–260 (2008)
[28] Than, C., Ruths, D., Innan, H., Nakhleh, L.: Identifiability issues in phylogeny-based detection of horizontal gene transfer. In: Bourque, G., El-Mabrouk, N. (eds.) RECOMB-CG 2006. LNCS (LNBI), vol. 4205, pp. 215–219. Springer, Heidelberg (2006)
[29] Than, C., Ruths, D., Innan, H., Nakhleh, L.: Confounding factors in HGT detection: Statistical error, coalescent effects, and multiple solutions. Journal of Computational Biology 14(4), 517–535 (2007)