Inferring Phylogenetic Networks by the Maximum Parsimony Criterion: A Case Study

Guohua Jin, Luay Nakhleh, Sagi Snir,† and Tamir Tuller‡*

*Department of Computer Science, Rice University, Houston, Texas; †Department of Mathematics, University of California; and ‡School of Computer Science, Tel Aviv University, Tel Aviv, Israel

Horizontal gene transfer (HGT) may result in genes whose evolutionary histories disagree with each other, as well as with the species tree. In this case, reconciling the species and gene trees results in a network of relationships, known as the "phylogenetic network" of the set of species. A phylogenetic network that incorporates HGT consists of an underlying species tree that captures vertical inheritance and a set of edges which model the "horizontal" transfer of genetic material. In a series of papers, Nakhleh and colleagues have recently formulated a maximum parsimony (MP) criterion for phylogenetic networks, provided an array of computationally efficient algorithms and heuristics for computing it, and demonstrated its plausibility on simulated data.

In this article, we study the performance and robustness of this criterion on biological data. Our findings indicate that MP is very promising when its application is extended to the domain of phylogenetic network reconstruction and HGT detection. In all cases we investigated, the MP criterion detected the correct number of HGT events required to map the evolutionary history of a gene data set onto the species phylogeny. Furthermore, our results indicate that the criterion is robust with respect to both incomplete taxon sampling and the use of different site substitution matrices. Finally, our results show that the MP criterion is very promising in detecting HGT in chimeric genes, whose evolutionary histories are a mix of vertical and horizontal evolution. Besides the performance analysis of MP, our findings offer new insights into the evolution of 4 biological data sets and new possible explanations of HGT scenarios in their evolutionary history.

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) (Lake et al. 1999; Eisen 2000a; Kurland 2000; Ochman et al. 2000; Jain et al. 2002, 2003; Brown 2003; Doolittle et al. 2003). Views as to the extent of HGT in bacteria vary between the 2 extremes (Doolittle 1999a, 1999b; Welch et al. 2002; Kurland et al. 2003; Hao and Golding 2004; McClilland et al. 2004; Nakamura et al. 2004). There is 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 merely "background noise" that does not affect the reconstructibility of a phylogenetic tree for bacterial genomes. Supporting arguments for these 2 views have been published. For example, the heterogeneity of genome composition between closely related strains (only 40% of the genes in common with 3 *Escherichia 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. Furthermore, recent evidence shows that HGT also plays a major evolutionary role in plants (Bergthorsson et al. 2003, 2004; Mower et al. 2004).

HGT is considered a primary explanation of incongruence among gene phylogenies and a significant obstacle to reconstructing the Tree of Life (Daubin et al. 2003). 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.

The goal of many biological studies has been to identify genes that were acquired by the organism through horizontal transfers rather than inherited from their ancestors. In one of the first papers on the topic, Medigue et al. (1991) proposed the use of multivariate analysis of codon usage to identify such genes; since then various authors have proposed other intrinsic methods, such as using GC content, particularly in the third position of codons (e.g., Lawrence and Ochman 1997). On the basis of such approaches, a database of putative horizontally transferred genes in prokaryotes has been established (Garcia-Vallve et al. 2003). An advantage of intrinsic approaches is their ability to identify and eliminate genes that do not obey a treelike process of evolution—genes that prevent classical phylogenetic methods from reconstructing an accurate tree. With the advent of whole-genome sequencing, more powerful intrinsic methods become possible, such as those using the location of suspect genes with each genome: such locations tend to be preserved through lineages, but a transfer event can place the new gene in a more or less random location. However, even advanced approaches are sensitive to differential selection pressures, uneven evolutionary rates, and biased sampling, all of which can give rise to false identification of HGT events (Eisen 2000b).

Nonintrinsic approaches use phylogenetic reconstructions to identify incongruence that can indicate transfer events (Daubin et al. 2003). Incongruence identification has been addressed with phylogenetic reconstruction as

Key words: reticulate evolution, phylogenetic networks, horizontal gene transfer, maximum parsimony, computational phylogenetics.

E-mail: nakhleh@cs.rice.edu.

Mol. Biol. Evol. 24(1):324–337. 2007
doi:10.1093/molbev/msl163
Advance Access publication October 31, 2006

follows: given DNA sequences for several genes, should the sequence data sets be combined and then analyzed or should they be analyzed separately and the analyses results reconciled? (e.g., Bull et al. 1993; Chippindale and Wiens 1994; Olmstead and Sweere 1994; de Queiroz et al. 1995; Huelsenbeck et al. 1996; Cunningham 1997; Wiens 1998). The standard conclusion that many genes inherited through lineal descent would override the confusing signal generated by a few genes acquired through horizontal transfer appears wrong (Teichmann and Mitchison 1999; Brown et al. 2001). Lawrence and Ochman (2002) surveyed some of these methods.

With whole-genome sequencing, extra information for resolving gene tree incongruence becomes available. Huynen and Bork (1998) advocate the use of 2 types of data: the fraction of shared orthologs and gene synteny. Synteny (the conservation of genes on the same chromosome) is not widely applicable to prokaryotes, but its logical extension, conservation of gene order, definitely is. Huynen and Bork proposed to measure the fraction of conserved adjacencies, a notion that had been introduced earlier by Sankoff in a series of papers defining break points (adjacencies that are not conserved) and their uses (Blanchette et al. 1997; Sankoff and Blanchette 1998). Proposing a different approach, Daubin et al. (2002) combined orthology search techniques and information from the DNA sequences themselves to improve the detection of horizontal transfers. Orthologs are a phylogenetic notion: 2 homologous genes are orthologs if they are the product of speciation from a common ancestor; in contrast, 2 homologous genes are paralogs if they are the product of duplication. However, determining orthologs can be difficult and has added to the complexity of the problem. Other methods, such as "quartet mapping," have been proposed recently, but they have been found to significantly overestimate the extent of HGT (Daubin and Ochman 2004). Finally, from a computational point of view, the problem can be formulated as a graph-theoretic problem of reconciling species and gene trees into phylogenetic networks (Moret et al. 2004). Computational approaches have been proposed by Hallett and Lagergren (Hallett and Lagergren 2001; Addario-Berry et al. 2003), Boc and Makarenkov (2003), and Nakhleh, Ruths, et al. (2005). A slightly different approach to the problem was taken by Kunin et al. (2005), in which they reconstructed the tree for "vertical inheritance" and used an ancestral state inference algorithm to map the HGT events to the tree, thus obtaining a network.

Maximum parsimony (MP) is one of the most popular methods used for phylogenetic tree reconstruction. Roughly this method is based on the assumption that "evolution is parsimonious," that is, the best evolutionary trees are the ones that minimize the number of changes along the edges of the tree. The tree sought is one that minimizes the total number of mutations along its branches. The MP criterion has been successfully used to study the evolution of various data sets for almost 30 years, and despite a heated debate concerning its performance, it is one of the most commonly used criteria for phylogeny reconstruction. In the early 1990s, Hein (1990, 1993) introduced an extension of the MP criterion to model the evolutionary history of a set of sequences in the presence of recombination. In 2005 and 2006, Nakhleh and colleagues gave a mathematical formulation of the MP criterion for phylogenetic networks and devised computationally efficient solutions aimed at reconstructing and evaluating the quality of phylogenetic networks under the MP criterion (Nakhleh, Jin, et al. 2005; Jin et al. 2006b). Furthermore, they investigated the performance of the criterion on small synthetic data sets.

In this article, we investigate the performance and robustness of the MP criterion for phylogenetic networks on real biological data sets. In particular, we study the performance of the MP criterion with respect to detecting the actual number and location of HGT events, the robustness of the criterion with respect to incomplete taxon sampling and different site substitution matrices, and the applicability of the criterion to detecting HGT in chimeric genes.

Our findings indicate that MP is very promising when extended to the domain of phylogenetic network reconstruction and HGT detection. In all cases we investigated, the MP criterion detected the correct number of HGT events required to map the evolutionary history of a gene data set onto the species phylogeny. Further, our results indicate that the criterion is robust with respect to both incomplete taxon sampling and the use of different site substitution matrices. Finally, our results show that the MP criterion is very promising in detecting HGT in chimeric genes, whose evolutionary histories are a mix of vertical and horizontal evolution.

Besides the performance analysis of MP, our findings offer new insights into the evolution of 4 biological data sets and new possible explanations of HGT scenarios in their evolutionary history. For the *rbcL* gene data set of (Delwiche and Palmer 1996), we identified 7 HGT edges, resolving some questions left open by the authors regarding the exact location of some of these edges. For the *rpl12e* gene data set, we identified 3 HGT edges, whose addition to the species tree explain its incongruence with the gene tree, as reported in Matte-Tailliez et al. (2002). In the case of the *rps11* gene data set of Bergthorsson et al. (2003), we identified 3 HGT edges, including one partial HGT involving the 3′ half of the gene in the *Sanguinaria* species. Finally, for the *cox2* gene data set of Bergthorsson et al. (2004), we identified 2 HGT edges, one which includes the only well-supported HGT postulated by the authors and another that is a reflection of the lack of resolution in the gene tree.

The rest of the paper is organized as follows. In the Materials and Methods, we briefly review the MP criterion for phylogenetic networks, describe the data sets we used, and explain the phylogenetic analyses we conducted and the questions we attempted to answer. In the Results and Discussion, we report on our findings and discuss the performance of the MP criterion with respect to 4 different questions. Finally, we review our recent introduction of the maximum likelihood (ML) criterion to phylogenetic networks (Jin et al. 2006a) and compare it with the parsimony criterion.

Materials and Methods

MP of Phylogenetic Networks

Phylogenetic networks model evolutionary histories in the presence of nontreelike events, such as HGT and hybrid

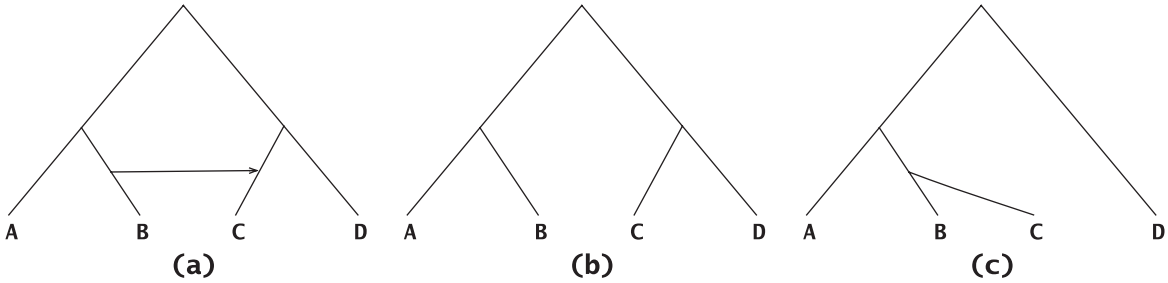


FIG. 1.—A phylogenetic network N on 4 taxa, A, B, C, and D, and the 2 trees it contains: $\mathcal{T}(N)=\{T_1, T_2\}$.

speciation. In the case of HGT, a phylogenetic network N is a rooted, directed, acyclic graph, whose leaves are labeled uniquely by a set of taxa and which consists of an underlying species tree augmented with a set of additional HGT edges. Additionally, the graph must satisfy certain temporal constraints; a formal description of the model and these constraints is given in Moret et al. (2004).

We say that a tree is "contained" in a phylogenetic network if it can be obtained from the network by the following 2 steps: 1) for every node in the network, remove all but one of the edges incident into it (i.e., the edges whose head is the node under consideration); and 2) for every node u with a single parent p and a single child c, remove u and the 2 edges incident to it and add a new edge from p to c (repeat this step as long as such nodes as u exist). Given a phylogenetic network N, we denote by $\mathcal{T}(N)$ the set of all trees contained inside N. Although a phylogenetic network models the evolutionary history of species, the evolution of each individual gene is modeled by some tree (except for cases we will handle later) contained in the network. Figure 1*a* shows a phylogenetic network on 4 taxa, with a single HGT edge that models horizontal transfer from species B to species C. The evolutionary history of the genes that evolve vertically is shown in figure 1*b*, whereas that of the horizontally transferred genes is shown in figure 1*c*.

This relationship between a phylogenetic network and its constituent trees is the basis for the MP extension to phylogenetic networks described. We now briefly review the definitions of Nakhleh, Jin, et al. (2005).

Definition 1. The Hamming distance between 2 equal-length sequences x and y, denoted by $H(x, y)$, is the number of positions j, such that $xj \neq yj$.

Given a fully labeled tree T, that is, a tree in which each node v is labeled by a sequence s_v over some alphabet Σ, we define the Hamming distance of an edge $e \in E(T)$, denoted by $H(e)$, to be $H(s_u, s_v)$, where u and v are the 2 endpoints of e. We now define the parsimony score of a tree T.

Definition 2. The parsimony score 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 score.

The parsimony definitions can be extended in a straightforward manner to incorporate different site substitution matrices, where different substitutions do not necessarily contribute equally to the parsimony score, by simply modifying the formula $H(x, y)$ to reflect the weights. Let Σ be the set of states that the 2 sequences x and y can take (e.g., $\Sigma = \{A, C, T, G\}$ for DNA sequences) and W the site substitution matrix such that $W[\sigma_1, \sigma_2]$ is the weight of replacing σ_1 by σ_2, for every σ_1, $\sigma \in \Sigma$. In particular, the "identity" site substitution matrix satisfies $W[\sigma_1, \sigma_2] = 0$ when $\sigma_1 = \sigma_2$, and $W[\sigma_1, \sigma_2] = 1$ otherwise. The weighted Hamming distance between 2 sequence is $H(x, y) = \sum_{1\leq i\leq k} W(x_i, y_i)$, where k is the length of the sequences x and y. The rest of the definitions are identical to the simple Hamming distance case.

Given a set S of sequences, the MP problem is to find an MP phylogenetic tree T for the set S. Unfortunately, this problem is nondeterministic polynomial (NP)-hard, even when the sequences are binary (Foulds and Graham 1982; Day 1983). One approach that is used in practice is to look at as many leaf-labeled trees as possible and choose one with a minimum parsimony score. The problem of computing the parsimony score of a fixed leaf-labeled tree is solvable in polynomial time (Fitch 1971; Hartigan 1973).

As described above, the evolutionary history of a single (nonrecombining) 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 score to fit a data set that evolved on a network is to define the parsimony score for each site as the minimum parsimony score of that site over all trees contained inside the network.

Definition 3 (Hein 1990, 1993; Nakhleh, Jin, et al. 2005). The parsimony score of a network N leaf labeled by a set S of taxa is

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

where $T\text{Cost}(T, s_i)$ is the parsimony score of site s_i on tree T.

This definition is illustrated in figures 2 and 3. 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 break point can be detected by experimenting with different block sizes.

Based on this criterion, we would want to reconstruct a phylogenetic network whose parsimony score is minimized.

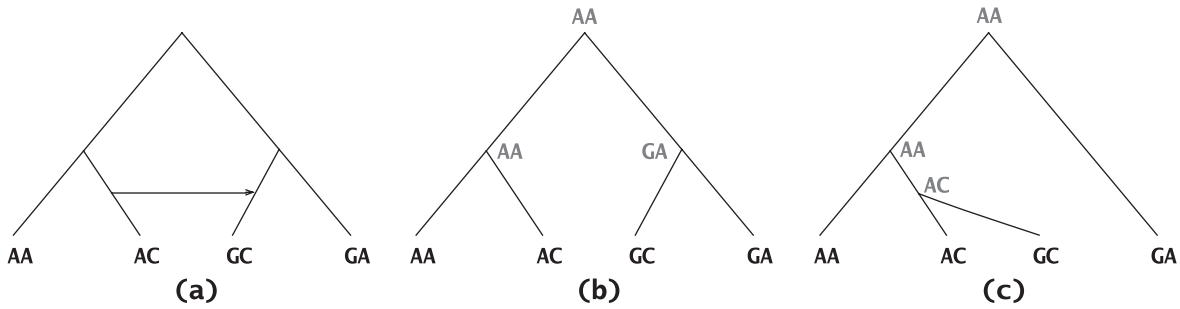


FIG. 2.—A phylogenetic network N on 4 taxa (*a*), each labeled by a sequence of length 2 so that there are 2 sites s_1 and s_2. An MP labeling of the internal nodes of the 2 trees T_1 (*b*) and T_2 (*c*) that are contained inside N are shown. $T\text{Cost}(T_1, s_1) = 1$, $T\text{Cost}(T_1, s_2) = 2$, $T\text{Cost}(T_2, s_1) = 2$, and $T\text{Cost}(T_2, s_2) = 1$. Based on Definition 3, $N\text{Cost}(N, S) = \min\{T\text{Cost}(T_1, s_1), T\text{Cost}(T_2, s_1)\} + \min\{T\text{Cost}(T_1, s_2), T\text{Cost}(T_2, s_2)\} = 1 + 1 = 2$. In this case, tree T_1 is the optimal tree for site s_1 and tree T_2 is the optimal tree for site s_2. In other words, under the maximum parsimony criterion, site s_1 evolved vertically under tree T_1 and site s_2 was horizontally transferred according to tree T_2. The phylogenetic network in figure 3 is optimally based on Definition 3.

In the case of HGT, a species tree that models vertical inheritance is usually known; for example, see Lerat et al. (2003). 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 HGT events. This is defined as the "fixed-tree MP on phylogenetic networks problem" in Nakhleh, Jin, et al. (2005).

Definition 4. Fixed-tree MP on phylogenetic networks (FTMPPN):

Input: A species tree T leaf labeled by a set S of sequences and a nonnegative integer k.
Output: A phylogenetic network N, consisting of T and a set X of additional HGT edges with $|X| = k$, which minimizes $N\text{Cost}(N, S)$.

A major challenge for solving the FTMPPN problem, as formulated in Definition 4, is that the value of k (the number of HGT edges to be added) is usually unknown and one of the outcomes sought by a biologist. This challenge is further complicated by the following observation.

Observation 1. Let N_1 and N_2 be 2 phylogenetic networks which are obtained by adding 2 sets X_1 and X_2 of HGT edges, respectively, to a species tree T, such that $X_1 \subseteq X_2$. Then,

1. $\mathcal{T}(N_1) \subseteq \mathcal{T}(N_2)$.
2. $N\text{Cost}(N_1, S) = N\text{Cost}(N_2, S)$.

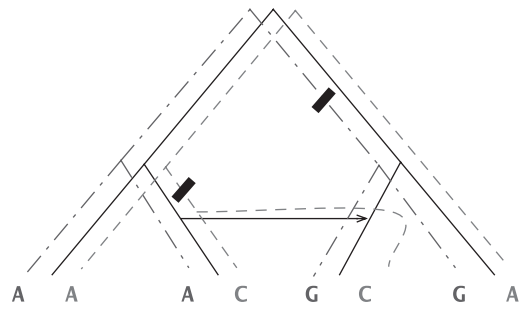


FIG. 3.—An MP phylogenetic network, with $N\text{Cost}(N, S) = 2$ for the leaf-labeled phylogenetic network in figure 2. The MP tree of the first site is described by the dash-dot lines, whereas the MP tree of the second site is described by the dashed lines. Each tree has a single mutation, and they are both reconciled inside the phylogenetic network with a single HGT edge, described by the solid lines.

The implication of Observation 1 is that as more edges are added to species tree T in solving the FTMPPN problem, the parsimony score either improves or remains unchanged but never gets worse. Therefore, we reformulate the FTMPPN problem so as to add more HGT edges as long as the improvement in the parsimony score is beyond a certain threshold.

Definition 5. Threshold-based FTMPPN (θ-FTMPPN):

Input: A species tree T leaf labeled by a set S of sequences and a positive threshold value θ.
Output: A phylogenetic network N consisting of T and a set X of HGT edges $h_1, h_2, \ldots, h_m$, such that

1. $|N\text{Cost}(N_{i+1}, S) - N\text{Cost}(N_i, S)| \geq \theta$, for $0 \leq i \leq m - 1$, where N_i, $i \geq 1$, is the phylogenetic network obtained by adding the HGT edges $h_1, h_2, \ldots, h_i$ to T, and $N_0 = T$;
2. the order in which the edges in X are added to T does not matter; and
3. X is maximal (set X is "maximal," if every set X', where $X \subset X'$, does not satisfy at least 1 of conditions (1) and (2) in the definition) among all sets of edges that satisfy (1) and (2).

Whereas it seems that the difference between Definition 4 and Definition 5 is merely obtained by "shifting the focus" in input parameter from k, the number of HGT edges to add, to θ, the threshold beyond which a parsimony improvement is considered significant, this shift is significant from a practical point of view. As we will show later, inspecting the parsimony improvement as more HGT edges are added, a clear "stopping rule" is determined for the most part (whereas such a rule cannot be determined based solely on the number of HGT edges added). In other words, θ plays the role of a parameter to control overfitting of the sequence data to the phylogenetic network.

A natural concern that Definition 5 raises is that the order in which the m HGT edges are added may affect the outcome and, hence, condition (2) in the definition. We conducted extensive studies to investigate this concern, and in all 4 data sets we analyzed, identical results were obtained regardless of the ordering of HGT edges that was employed.

Therefore, shifting the focus from the number of HGT edges required to a threshold of improvement significance, coupled with the empirical observation that the order in which HGT edges are added does not affect the final outcome, substantiates the practical applicability of Definition 5.

Nakhleh, Jin, et al. (2005) provided an exhaustive solution for the FTMPPN. Their algorithm is based on the empirical evidence that an MP network with k HGT edges is obtained by adding an HGT edge to an optimal network with $k - 1$ edges (the experiments were conducted several times, while taking the HGT edges in different orders, and in all these cases, the same resulting networks were obtained). The algorithm seeks to add an edge in all possible ways to all optimal networks with $k - 1$ network edges. This approach requires computing the parsimony score, based on Definition 3, of every phylogenetic network obtained during the search; we refer to this computation as the parsimony score of phylogenetic networks (PSPN) problem.

Definition 6. PSPN:

Input: A set S of aligned sequences and a phylogenetic network N leaf labeled by S.
Output: NCost(N, S).

Nakhleh et al. provided a straightforward algorithm for solving the PSPN problem, which enumerates all trees contained inside the network and therefore runs in O(mln) time, where $m=|T(N)|$, l is the sequence length, and n is the number of taxa (leaves) in the phylogenetic network. Because the number of HGT edges is O(n^2) in the worst case, the number of trees inside a network may be exponential in the number of leaves, and hence, the running time of the algorithm is exponential in the number of HGT edges (in the number of taxa in the worst case). We proved that the PSPN problem is NP-hard and developed more computationally efficient algorithms and heuristics for the PSPN and FTMPPN problems in Jin et al. (2006b).

Data Sets

We analyzed 4 biological data sets with the aim of identifying the number of HGT events, as well as their respective donors and recipients:

1. The rubisco gene *rbcL* of a group of 46 plastids, cyanobacteria, and proteobacteria, which was analyzed by Delwiche and Palmer (1996). This data set consists of 46 aligned amino acid sequences (each of length 532), 40 of which are from form I of rubisco and the other 6 are from form II of rubisco. The first 21 and the last 14 sites of the sequence alignment were excluded from the analysis, as recommended by the authors. The species tree for the data set was created based on information from the ribosomal database project (http://rdp.cme.msu.edu) and the work of Delwiche and Palmer (1996).
2. The ribosomal protein *rpl12e* of a group of 14 archaeal organisms, which was analyzed by Matte-Tailliez et al. (2002). This data set consists of 14 aligned amino acid sequences, each of length 89 sites. The authors constructed the species tree using ML, once on the concatenation of 57 ribosomal proteins (7,175 sites) and another on the concatenation of small- and large-subunit rRNA (3,933 sites). The 2 trees are identical, except for the resolution of the *Pyrococcus* 3-species group; we used the tree based on the ribosomal proteins.
3. The ribosomal protein gene *rps11* of a group of 47 flowering plants, which was analyzed by Bergthorsson et al. (2003). This data set consists of 47 aligned DNA sequences, each with 456 sites. The authors analyzed the 3′ end of the sequences separately; this part of the sequences contains 237 sites. The species tree was reconstructed based on various sources, including the work of Michelangeli et al. (2003) and Judd and Olmstead (2004).
4. The mitochondrial gene *cox2* of a group of 25 seed and nonseed plants, which was analyzed by Bergthorsson et al. (2004). This data set consists of 28 aligned DNA sequences including 4 copies of the *Amborella* gene. Each aligned sequence is 311 bp long. Ten regions including primer sites and editing sites were excluded from the analysis, as suggested by the authors. The authors generated an MP tree from which an ML tree was built based on estimated parameters. The ML tree was further refined into a stable state. Seed and nonseed plants were analyzed separately. We used a species tree for the data set based on information at National Center for Biotechnology Information (http://www.ncbi.nih.gov) and analyzed the entire data set with both seed and nonseed plants.

Phylogenetic Analyses

To understand the performance of MP as a criterion for reconstructing phylogenetic networks, in general, and detecting HGT, in particular, we investigated the 4 data sets with respect to 4 different questions.

1. Does the MP criterion correctly identify the number and location (i.e., donors and recipients) of HGT events needed to explain the evolutionary history of a gene with respect to a species phylogeny? To answer this question, we analyzed the *rbcL*, *rpl12e*, and *cox2* data sets by running the methods of Jin et al. (2006b) for solving the FTMPPN problem. More specifically, for each of the data sets, we sought a set of edges whose addition to the species tree yielded an optimal phylogenetic network under the parsimony criterion. As discussed before, adding more edges to a phylogenetic network either improves the parsimony score of the phylogenetic network or leaves it unchanged. We analyzed the rate of improvement as a way to determine when to stop adding extra edges. The quality of the MP criterion with respect to this question (as well as the next 3 questions) was determined by comparing our findings to the hypotheses postulated by the authors of the data sets we considered.
2. How does incomplete taxon sampling affect the performance of the criterion with respect to question (1)? To answer this question, we sampled 15 taxa from the *rbcL* data set in such a way to ensure that the donors and recipients of the HGTs detected in the analysis of question (1) were present among these 15 taxa. We used this 15-taxon data set to study the performance of the MP

criterion with respect to detecting the number of HGTs as well as the donor/recipient of these HGTs.
3. Does the site substitution matrix affect the performance of the criterion? To answer this question, we reanalyzed the 15-taxon *rbcL* data set under various amino acid substitution matrices. Once again, we analyzed the data set with respect to the number as well as location of the HGTs.
4. Can the MP criterion help identify partial (chimeric) HGT? Bergthorsson et al. (2003) analyzed the *rps11* gene in a group of flowering plants and postulated that not only did this gene involve HGT but also that it was chimeric: its 5′ half was vertically inherited, whereas its 3′ half was horizontally transferred. We analyzed both the complete *rps11* gene as well as its 3′ half.

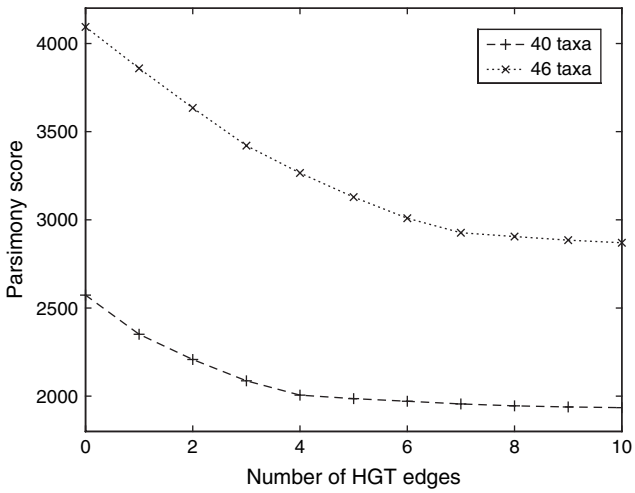


FIG. 4.—Optimal improvement in the parsimony score as extra edges are added to the species tree to obtain a phylogenetic network on the *rbcL* data set. The most significant improvements are obtained by adding the first 7 HGT edges in the case of the 46-taxon data set and the first 4 HGT edges in the case of the 40-taxon data set.

Results and Discussion

Identifying the Numbers and Locations of HGT Events

In this section, we report on our findings when analyzing the *rbcL*, *rpl12e*, and *cox2* data sets using our methods for solving the FTMPPN. For each data set we show the optimal improvement in parsimony scores as new edges are added to the species trees, as well as the location of the edges that correspond to these optimal improvements. All results in this section are based on the identity substitution matrix (which assigns value 0 to 2 identical sites and value 1 to 2 different sites).

The *rbcL* Gene Data Set

We analyzed the *rbcL* gene data set twice in this context: once with the complete data set of 46 organisms and another with only 40 organisms; the latter was obtained by removing the form II rubisco. (The original data set had 48 species, but the 2 species endosymbiont of *Alvinoconcha* and *Pseudomonas hydrogenothermophila* were unclassified in Delwiche and Palmer (1996); hence, we excluded them). The optimal improvements in the parsimony score as 10 edges are added to both species trees of this data set are shown in figure 4.

We observe that the optimal improvement in parsimony score is always higher than 80 points for every edge added of the first 7 in the case of the 46-taxon data set and the first 4 in the case of the 40-taxon data set. The actual HGT edges that correspond to the first 7 and first 4 of these optimal improvements for the 46- and 40-taxon data sets are shown in figure 5*a* and *b*, respectively.

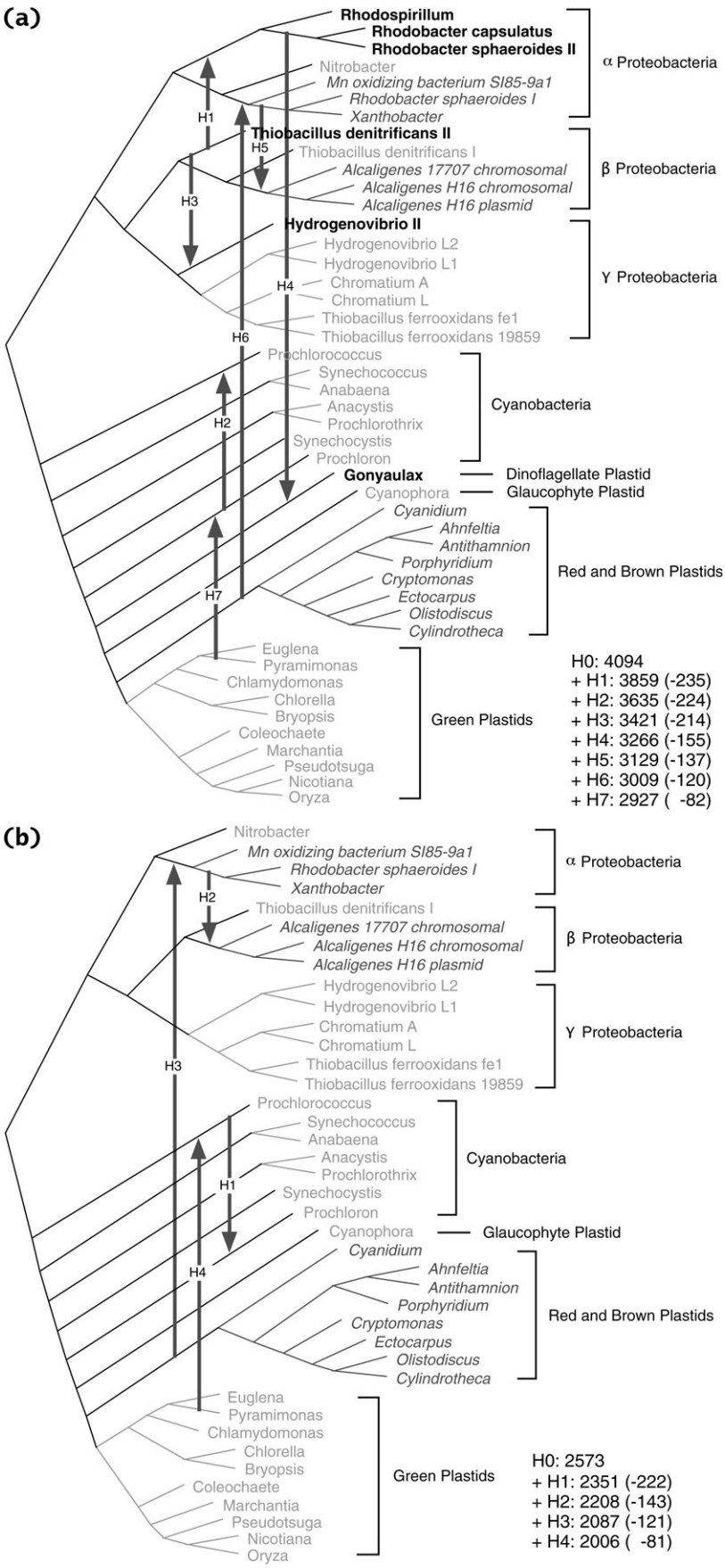
The HGT edges H1, H3, and H4 in figure 5*a* group all the form II species together; because these species are excluded from the 40-taxon data set, these edges have no equivalent ones in figure 5*b*. The remaining 4 HGT edges, H2, H5, H6, and H7, in figure 5*a* achieve the same effect of the 4 HGT edges H1, H2, H3, and H4 in figure 5*b*. Edges H2 and H5 in figure 5*b* and *a*, respectively, group the 3 *Alcaligenes* species together with the *Rhodobacter sphaeroides I* and *Xanthobacter* species, indicating an HGT from the most recent common ancestor of the latter 2 species to the most recent common ancestor of the *Alcaligenes* species. Edges H3 and H6 in figure 5*b* and *a*, respectively, group 3 α-proteobacteria with the group of red and brown plastids. Edges H1 and H4 in figure 5*b* and edges H2 and H7 in figure 5*a* indicate different HGTs in the 2 data sets, yet achieve exactly the same grouping: they group the 2 cyanobacteria *Prochlorococcus* and *Prochloron* together and then group these 2 together with the green plastid *Pyramimonas*.

How do these findings compare with the hypotheses of Delwiche and Palmer (1996)? The authors postulated that at least 4 independent HGTs were required to explain the division of plastids and proteobacteria into the greenlike and redlike groups:

> A transfer of redlike rubisco operon from a proteobacterium to a common ancestor of red and brown plastids. Our analysis computed such a transfer, but we found that it is from a common ancestor of red and brown plastids to a proteobacterium (edge H6 in fig. 5*a*).
>
> A transfer of a cyanobacterial greenlike *rbcL* to an ancestor of γ-proteobacteria early in their evolution (but after the emergence of the β-proteobacteria from within the γ-proteobacteria).
>
> A transfer from this same γ-proteobacterial lineage to ancestor of the α-proteobacterium *Nitrobacter vulgaris*.
>
> A transfer from this same γ-proteobacterial lineage to ancestor of the β-proteobacterium *Thiobacillus denitrificans*.

In the case of the last 3 transfers, the authors were not certain about them (even postulating that the incongruence may be due to inaccurate identification of some of the taxa). Our analysis, instead, indicates 2 transfers from the β-proteobacterium *Thiobacillus denitrificans II* to the α and γ groups of the form II rubisco. Furthermore, they postulated 3 more HGTs to account for incongruities in the *rbcL* phylogeny:

> A transfer of the *Gonyaulax* rubisco (in the gene tree, it is grouped within the α-proteobacteria of the form II


(a)
Rhodospirillum
Rhodobacter capsulatus
Rhodobacter sphaeroides II
Nitrobacter
Mn oxidizing bacterium SI85-9a1
Rhodobacter sphaeroides I
Xanthobacter
α Proteobacteria
Thiobacillus denitrificans II
Thiobacillus denitrificans I
Alcaligenes 17707 chromosomal
Alcaligenes H16 chromosomal
Alcaligenes H16 plasmid
β Proteobacteria
Hydrogenovibrio II
Hydrogenovibrio L2
Hydrogenovibrio L1
Chromatium A
Chromatium L
Thiobacillus ferrooxidans fe1
Thiobacillus ferrooxidans 19859
γ Proteobacteria
Prochlorococcus
Synechococcus
Anabaena
Anacystis
Prochlorothrix
Synechocystis
Prochloron
Cyanobacteria
Gonyaulax
Dinoflagellate Plastid
Cyanophora
Glaucophyte Plastid
Cyanidium
Ahnfeltia
Antithamnion
Porphyridium
Cryptomonas
Ectocarpus
Olistodiscus
Cylindrotheca
Red and Brown Plastids
Euglena
Pyramimonas
Chlamydomonas
Chlorella
Bryopsis
Coleochaete
Marchantia
Pseudotsuga
Nicotiana
Oryza
Green Plastids
H1
H2
H3
H4
H5
H6
H7
H0: 4094
+ H1: 3859 (-235)
+ H2: 3635 (-224)
+ H3: 3421 (-214)
+ H4: 3266 (-155)
+ H5: 3129 (-137)
+ H6: 3009 (-120)
+ H7: 2927 (-82)
(b)
Nitrobacter
Mn oxidizing bacterium SI85-9a1
Rhodobacter sphaeroides I
Xanthobacter
α Proteobacteria
Thiobacillus denitrificans I
Alcaligenes 17707 chromosomal
Alcaligenes H16 chromosomal
Alcaligenes H16 plasmid
β Proteobacteria
Hydrogenovibrio L2
Hydrogenovibrio L1
Chromatium A
Chromatium L
Thiobacillus ferrooxidans fe1
Thiobacillus ferrooxidans 19859
γ Proteobacteria
Prochlorococcus
Synechococcus
Anabaena
Anacystis
Prochlorothrix
Synechocystis
Prochloron
Cyanobacteria
Cyanophora
Glaucophyte Plastid
Cyanidium
Ahnfeltia
Antithamnion
Porphyridium
Cryptomonas
Ectocarpus
Olistodiscus
Cylindrotheca
Red and Brown Plastids
Euglena
Pyramimonas
Chlamydomonas
Chlorella
Bryopsis
Coleochaete
Marchantia
Pseudotsuga
Nicotiana
Oryza
Green Plastids
H1
H2
H3
H4
H0: 2573
+ H1: 2351 (-222)
+ H2: 2208 (-143)
+ H3: 2087 (-121)
+ H4: 2006 (-81)

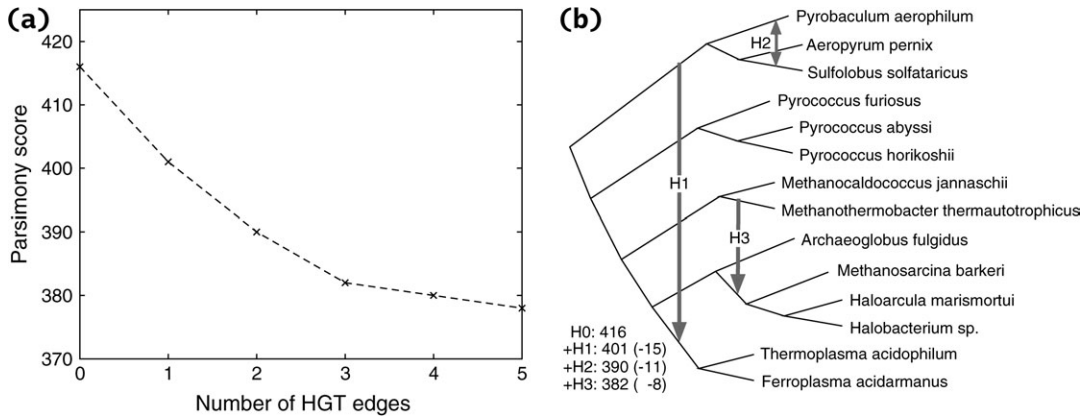


FIG. 6.—Results of the analysis on the *rpl12e* gene data set. (*a*) The most significant decrease in the parsimony score of the phylogenetic network as more HGT edges are added. The significant improvement was achieved after adding the first 3 HGT edges. (*b*) The phylogenetic network with the 3 HGT edges that resulted in the most significant improvement in the parsimony score. The improvement in the parsimony score per HGT edge is shown next to the phylogenetic network. "H0: X" indicates that X is the parsimony score of the species tree. "+H*i*: X (Y)" indicates that the parsimony score of the phylogenetic network after adding the *i*th HGT edge is X, and the decrease in the parsimony score achieved by HGT edge H*i* alone is Y.

rubisco). Our analysis indicates an HGT from the common ancestor of *Rhodobacter capsulatus*/*Rhodobacter sphaeroides II* to *Gonyaulax* (edge H4 in fig. 5*a*), which results in a grouping identical to that based on the *rbcL* gene tree in Delwiche and Palmer (1996).

A transfer from the greenlike proteobacterial group to *Prochlorococcus*. In this case, the authors could not determine with certainty where the transfer occurred. Our analysis shows that the transfer occurred from the cyanobacterium *Prochloron* to the cyanobacterium *Prochlorococcus* (edge H2 in fig. 5*a*).

A transfer involving one of the 3 groups: *Rhodobacter*/*Xanthobacter*, *Alcaligenes*, and Mn-oxidizing bacterium. In this case as well, the authors could not determine with certainty which of the 3 groups involved the transfer. Our analysis shows that the transfer occurred from the *Rhodobacter*/*Xanthobacter* group to the *Alcaligenes* group (edge H5 in fig. 5*a*).

Finally, our analysis gave rise to edge H7 in figure 5*a*, which gives indication of a transfer that was not postulated by the authors, but among all 7 edges found in our analysis, this edge led to the smallest improvement in the parsimony score.

Edges H1, H3, and H4 in figure 5*a* all correspond to form II rubisco; since these taxa are not present in the 40-taxon species tree, only the 4 remaining HGT edges were identified, and they are shown in figure 5*b*—the correspondence is described above.

The *rpl12e* Gene Data Set

Our analysis of the data set of (Matte-Tailliez et al. 2002) inferred 3 significant HGT edges, shown in figure 6. Edge H1 resulted in the most significant improvement in the parsimony score and indicated the transfer which was postulated by the authors. Edge H2 accounts for the incongruence between the species tree and the tree based on the *rpl12e* protein, where the 2 trees differ in the phylogenetic pattern of the *Aeropyrum pernix*/*Pyrobaculum aerophilum*/*Sulfolobus solfataricus* group. Edge H3 indicates a transfer between *Methanobacterium thermoautotrophicum* and the group of *Methanosarcina barkeri*, *Haloarcula marismortui*, and *Halobacterium* sp.

The *cox2* Gene Data Set

We identified 2 HGT edges using the MP criterion, as shown in figure 7. The most significant improvement in the parsimony score came from a horizontal transfer between 2 *Magnoliid* species, *Asarum* and *Laurus*. An equally significant improvement in the parsimony score was due to a transfer to *Amborella* from any one of the 3 mosses, namely, *Thuidium*, *Hypnum*, and *Brachythecium*, or an ancestor of these 3 mosses. This identification of horizontal transfer from a moss donor to *Amborella* matches very well with the results of Bergthorsson et al. (2004), who mainly studied HGTs to *Amborella*. Bergthorsson et al. postulated 3 HGTs to *Amborella* and used the SH test procedure (Shimodaira and Hasegawa 1999) to estimate the support of these 3 events. These 3 HGT edges consisted of 1 from a moss donor, with the strongest evidence and support from the SH test (<0.001) and bootstrap value ($>90\%$), and 2 from other angiosperms, which the authors deemed insignificant based on the SH test (>0.05) and weak bootstrap supports due to largely poorly resolved trees within angiosperms. As noted, the HGT edge H1, which was found to be most significant under the parsimony criterion, corresponds to the

←

FIG. 5.—The MP phylogenetic networks of the *rbcL* data set obtained by adding 7 edges for the 46-taxon case (*a*) and 4 edges for the 40-taxon case (*b*) to the underlying species tree. Parsimony score improvement after incrementally adding the HGTs are shown at the bottom. Form II rubisco are shown in bold. The improvement in the parsimony score per HGT edge is shown next to each phylogenetic network. "H0: X" indicates that X is the parsimony score of the species tree. "+H*i*: X (Y)" indicates that the parsimony score of the phylogenetic network after adding the *i*th HGT edge is X, and the decrease in the parsimony score achieved by HGT edge H*i* alone is Y.

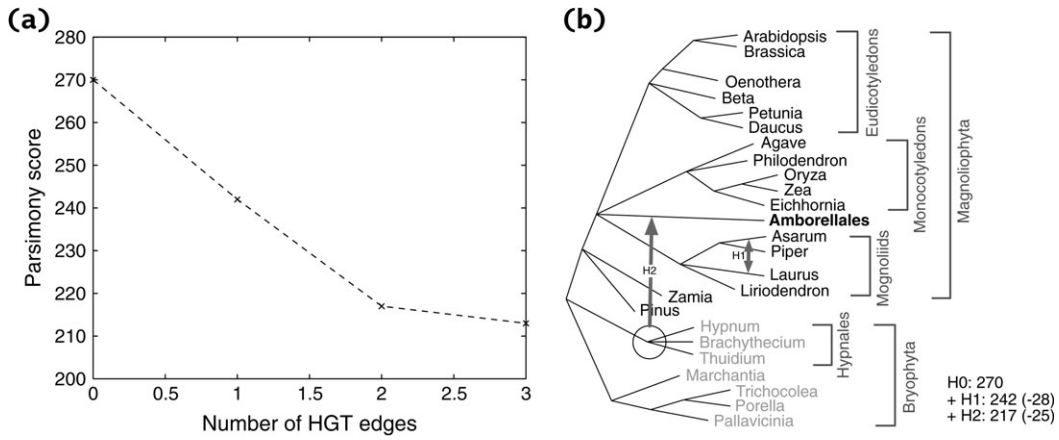


FIG. 7.—Results of the analysis on the *cox2* gene data set. (*a*) The most significant decrease in the parsimony score of the phylogenetic network as more HGT edges are added. The significant improvement was achieved after adding the first 2 HGT edges. (*b*) The phylogenetic network with the 2 HGT edges that resulted in the most significant improvement in the parsimony score. The circle at the source of the second HGT edge denotes that any of the tree edges within the circle could be the donor of the HGT, and with the same effect on the parsimony score. The improvement in the parsimony score per HGT edge is shown next to the phylogenetic network. "H0: X" indicates that X is the parsimony score of the species tree. "+H*i*: X(Y)" indicates that the parsimony score of the phylogenetic network after adding the *i*th HGT edge is X, and the decrease in the parsimony score achieved by HGT edge H*i* alone is Y.

only well-supported HGT event that Bergthorsson et al. found. The second HGT edge, H2, that was found by the parsimony analysis is probably a reflection of the "weak" phylogenetic signal in these gene sequence data (which is reflected by the weak support of most branches in the gene tree of *cox2* in Bergthorsson et al. [2004]).

Effects of Incomplete Taxon Sampling

To investigate the effects of incomplete taxon sampling on the performance of the MP criterion for detecting HGT, we selected 15 taxa from the *rbcL* data set so as to cover all the groups in the data set. These 15 taxa are shown at the tips of the tree in figure 8*b*. The improvement in parsimony score as extra edges are added to the species tree as well as the edges that correspond to the optimal improvements are shown in figure 8*a* and *b*, respectively. The figure shows clearly that the first 5 added edges lead to significant improvement in the parsimony score of the resulting phylogenetic network, whereas the improvement afterward is relatively much less significant.

Our results indicate that in this case we observe a similar trend to that of the full data set in terms of the improvement in parsimony score, yet slightly different results in terms of the locations of the HGT edges themselves. Since *Prochloron*—the donor in the HGT event H2 in figure 5*a*—was not sampled, the analysis did not detect the HGT event that involved it. Edges H1, H3, and H4 in figure 5*a* all involved form II species. Their counterparts in the analysis of the 15-taxon data set are edges H1, H2, and

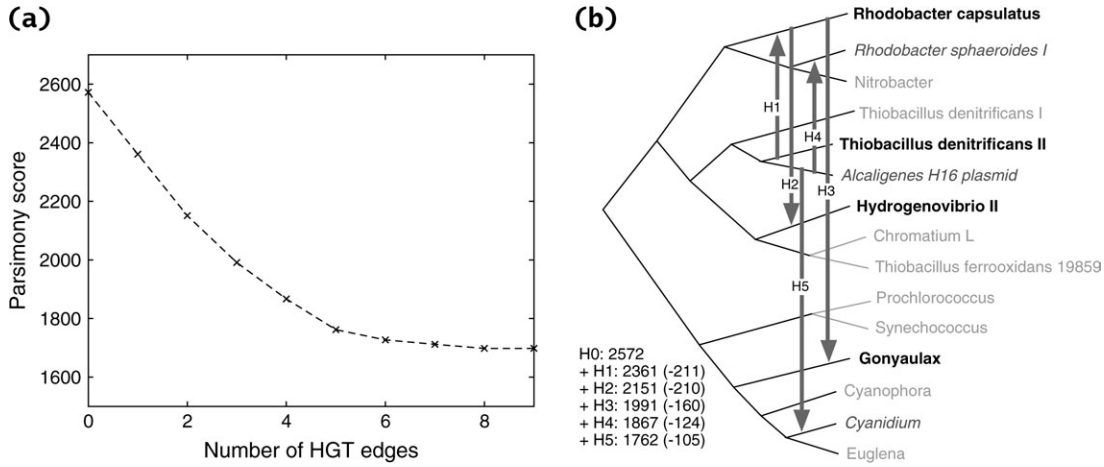


FIG. 8.—Results of the analysis on the 15-taxon *rbcL* gene data set. (*a*) The most significant decrease in the parsimony score of the phylogenetic network as more HGT edges are added. The significant improvement in the parsimony score was achieved after adding the first 5 HGT edges. (*b*) The phylogenetic network with the 3 HGT edges that resulted in the most significant improvement in the parsimony score. The improvement in the parsimony score per HGT edge is shown next to the phylogenetic network. "H0: X" indicates that X is the parsimony score of the species tree. "+H*i*: X (Y)" indicates that the parsimony score of the phylogenetic network after adding the *i*th HGT edge is X, and the decrease in the parsimony score achieved by HGT edge H*i* alone is Y.

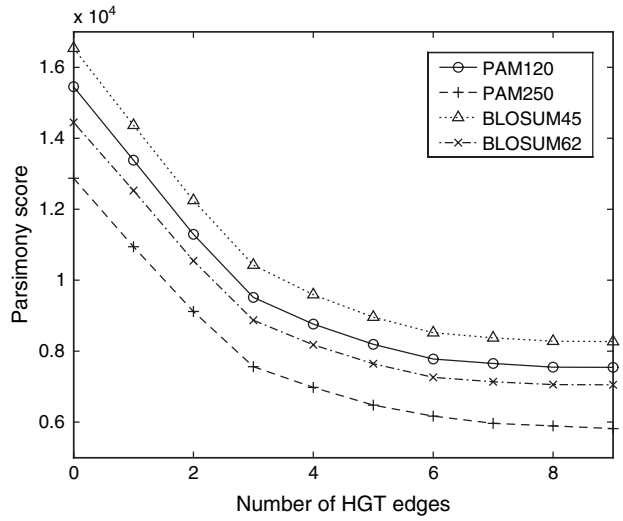


FIG. 9.—The most significant decrease in the parsimony score of the phylogenetic network as more HGT edges are added in the analysis of the 15-taxon *rbcL* data set under 4 different site substitution matrices. In all 4 cases, the significant improvement in the parsimony score was achieved after adding the first 6 HGT edges.

H3, respectively, as shown in figure 8*b*. Notice that these edges in both figures achieve the same result, namely, the grouping of the form II species, yet they differ in explaining the donor of the *rbcL* gene in the *Hydrogenovibrio II* species (edges H3 and H2 in figs. 5*a* and 8*b*, respectively).

Edges H4 and H5 in figure 8*b* achieve the same grouping as edges H5 and H6 in figure 5*a*, yet they place the groups in different places in their respective trees. Notice that since neither the donor nor the recipient of HGT edge H7 in figure 5*a* is present in the 15-taxon data set, no counterpart to this edge was found when analyzing the 15-taxon data set.

To conclude about the findings in this analysis, incomplete taxon sampling does not seem to affect the trend in parsimony score improvement, whereas it may have an impact on the directions of the HGT edges detected, as illustrated in the differences between the phylogenetic networks of figures 5*a* and 8*b*. A very significant implication of the results of this analysis is that the MP criterion is robust with respect to incomplete taxon sampling in terms of identifying the number of HGT events, as well the identity of the donors and recipients, yet it may get the directions of the HGT edges in reversed order.

Effects of Site Substitution Matrix

To investigate the robustness of the MP criterion for HGT detection with respect to the site substitution matrix used in the analysis, we reanalyzed the 15-taxon *rbcL* data set using 4 different matrices, in addition to the identity matrix used in the previous section: BLOSUM 45, BLOSUM 62, PAM 250, and PAM 120. The improvements in parsimony scores as extra HGT edges are added are shown in figure 9, and the phylogenetic networks with the HGT edges resulting in the optimal improvements are shown in figure 10.

In terms of the improvement in parsimony scores as extra HGT edges are added, figure 9 shows trends similar to that when using the identity matrix, as shown in figure 8*a*. The only difference is that in the case of these 4 matrices, the trends indicate 6 extra HGT edges, rather than 5 edges, as in the case of the identity matrix.

As for the detected HGT edges themselves, the results are very similar to those under the identity matrix in terms of the species grouping they achieve. The HGT edges detected under both BLOSUM matrices were identical, whereas these were different from the edges detected under the 2 PAM matrices, which also differed between them, as shown in the 3 phylogenetic networks in figure 10. The 3

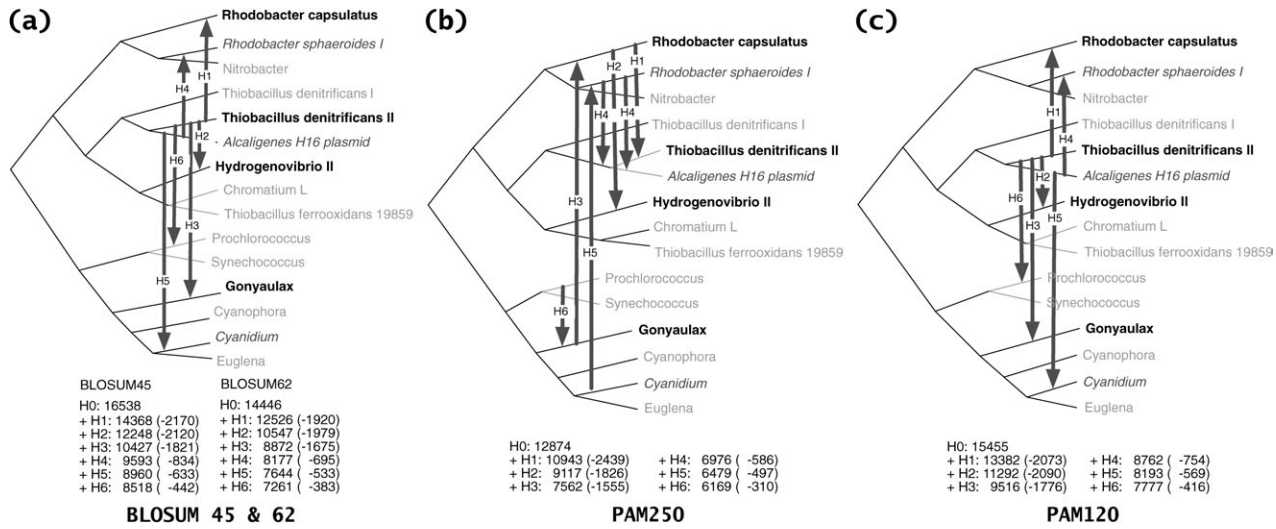


FIG. 10.—The MP phylogenetic networks of the 15-taxon *rbcL* data set obtained with the PAM and BLOSUM matrices. The improvement in the parsimony score per HGT edge is shown next to each phylogenetic network. "H0: X" indicates that X is the parsimony score of the species tree. "+H*i*: X (Y)" indicates that the parsimony score of the phylogenetic network after adding the *i*th HGT edge is X, and the decrease in the parsimony score achieved by HGT edge H*i* alone is Y. Having more than one HGT edge H*i* in a phylogenetic network indicates that this edge can be added in any of these locations, yet leading to exactly the same improvement in the parsimony score.

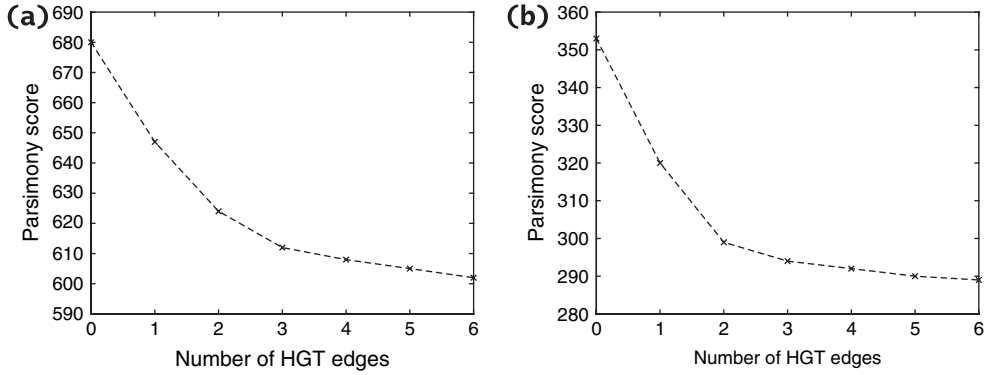


FIG. 11.—The most significant decrease in the parsimony score of the phylogenetic network as more HGT edges are added for the *rps11* data set. (*a*) The complete *rps11* sequences. (*b*) The 3′ end of the *rps11* sequences. In both cases, the significant improvement in the parsimony score was achieved after adding the first 3 HGT edges.

edges H1, H2, and H3 in all 3 phylogenetic networks have exactly the same effect, namely, the grouping of all form II species. Nonetheless, the phylogenetic networks differ in the placement of the groups. Edges H4 and H5 achieve the same result as that achieved by H4 and H5 under the identity matrix. In conclusion, the first 5 HGT edges detected under the various site substitution matrices achieve the same results. The sixth edge detected under the 4 matrices, but not the identity matrix, involves *Prochlorococcus*. This edge had the least significant contribution to improving the parsimony score among all 6 edges.

Detection of Partial HGT

In their analysis of a group of flowering plants, Bergthorsson et al. (2003) reported on transfers that created

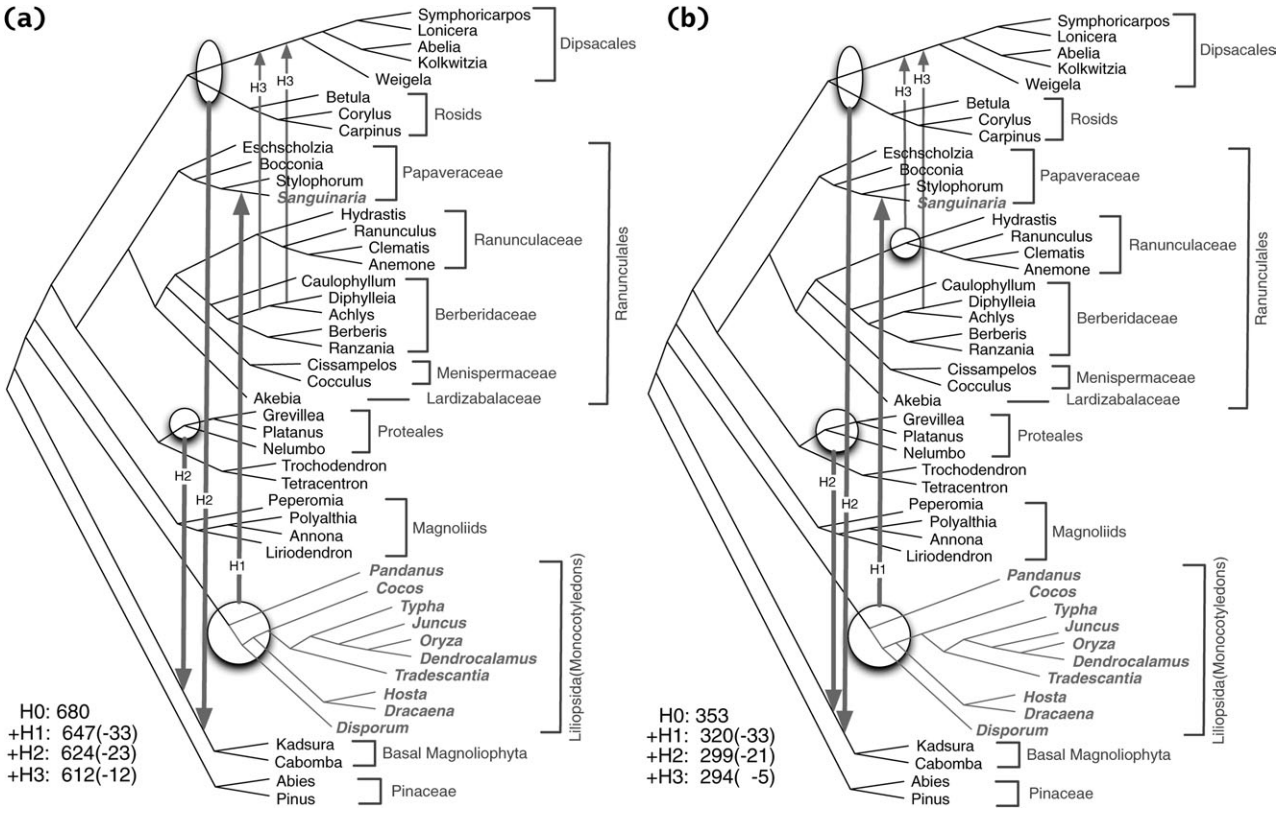


FIG. 12.—The phylogenetic network with the 3 HGT edges that resulted in the most significant improvement in the parsimony score. (*a*) The complete *rps11* sequences. (*b*) The 3′ end of the *rps11* sequences. The circles at the source of an HGT edge denote that any of the tree edges within that circle could be the donor of that respective HGT, and with the same effect on the parsimony score. For example, the donor of the HGT event denoted by edge H1 can be any of the 7 tree edges within the circle, all of which contribute equally to the improvement in the parsimony score. Having more than one HGT edge H*i* in a phylogenetic network indicates that this edge can be added in any of these locations, yet leading to exactly the same improvement in the parsimony score. The improvement in the parsimony score per HGT edge is shown next to each phylogenetic network. "H0: X" indicates that X is the parsimony score of the species tree. "+H*i*: X (Y)" indicates that the parsimony score of the phylogenetic network after adding the *i*th HGT edge is X, and the decrease in the parsimony score achieved by HGT edge H*i* alone is Y.

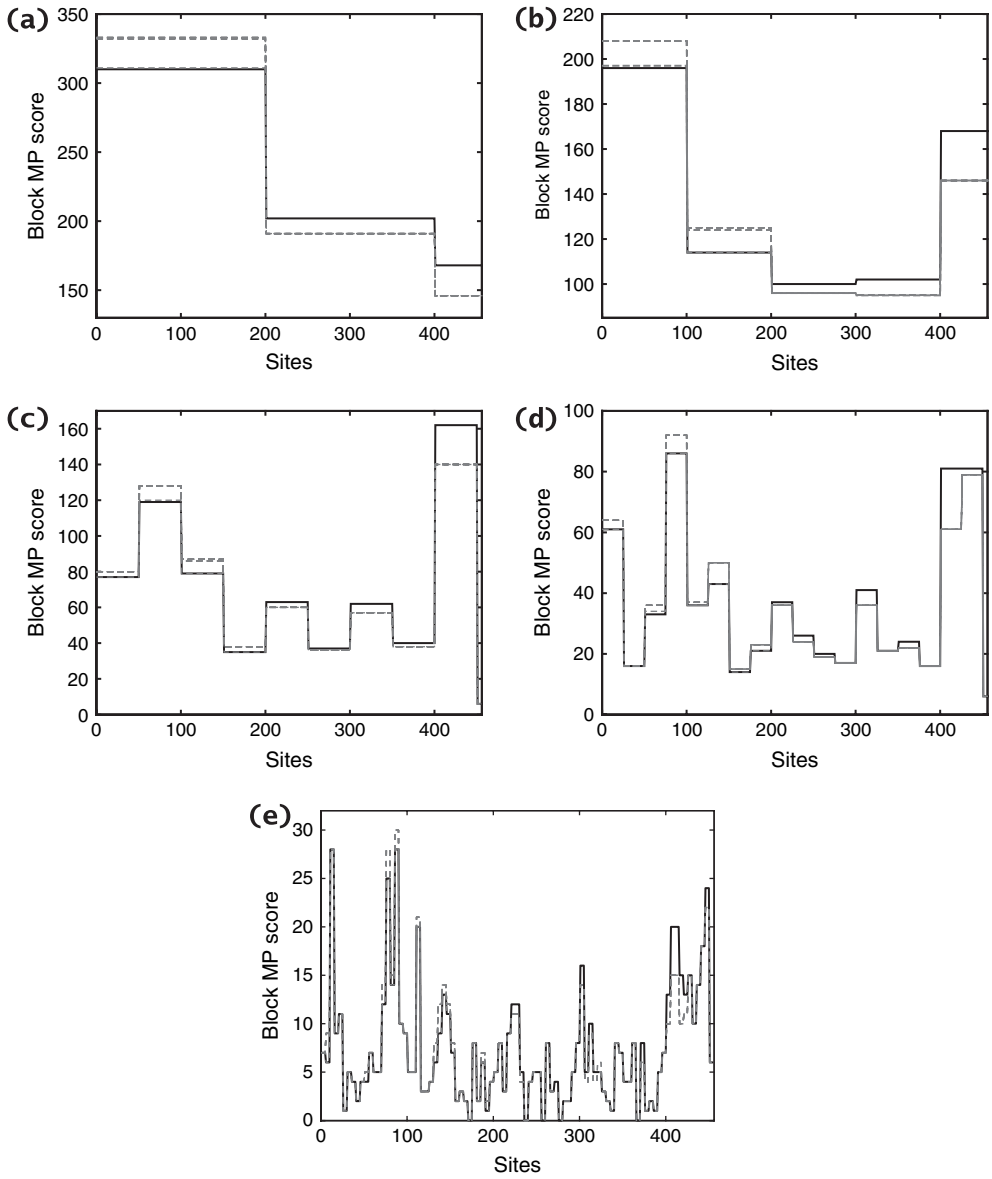


FIG. 13.—Blockwise analysis of the parsimony scores of the sequences on the species tree (solid line) and the 7 gene trees obtained by using the HGT edge H1 in figure 12 (dashed line). There are7 possible sources for H1, and hence the 7 possible gene trees. The blocks used in (*a*)–(*e*) consist of 200, 100, 50, 25, and 5 sites, respectively.

chimeric, half-monocot, half-dicot genes. In particular, they showed that *Sanguinaria rps11* is chimeric: its 5′ half is of expected eudicot, vertical origin, whereas its 3′ half is "indisputably of monocot, horizontal origin."

To investigate the performance of the MP criterion for detection HGT events in the case of chimeric genes, we analyzed the complete *rps11* gene sequences as well as their 3′ half separately. The improvements in the parsimony scores as HGT edges are added are shown in figure 11, and the phylogenetic networks themselves are shown in figure 12. The parsimony score improvement has almost identical trends in both cases of the complete and partial gene data sets, where in both cases it indicates that at most 3 HGT edges need to be added.

A significant observation is that in both cases, the first 2 HGT edges detected in the analysis are identical. Furthermore, the first edge, H1, leads to exactly the same improvement in the parsimony score (the parsimony score drops by 33 points in both cases). This clearly indicates that all the improvement in the parsimony score results in a transfer that involves only the 3′ half of the *rps11* gene. Similarly, edge H2 leads to almost the same drop in the parsimony score in both cases: 23 points in the case of the complete gene and 21 points in the case of its 3′ half, once again indicating the transfer of the 3′ half only.

In this analysis, we used the knowledge that the 3′ half was involved in the transfer and hence we were able to conduct the analysis on the complete gene as well as on its 3′ half. An interesting question is whether the MP criterion could detect that the 3′ half was involved in the transfer without having the knowledge a priori. To answer this question, we considered a phylogenetic subnetwork obtained from the

phylogenetic network in figure 12*a*, by taking only the species tree and the HGT edge H1, which is the one that corresponds to HGT in *Sanguinaria*. Because there are 7 possible donors for the HGT denoted by H1, this phylogenetic network in fact represents 7 networks, each of which contains the species tree and the gene tree obtained by moving the *Sanguinaria* close to the Monocotyledons. For each such phylogenetic network, we passed a window of a fixed size across the sequence alignment and computed the parsimony score of the block within the window on the species as well as the 7 gene trees inside the network. Figure 13 shows the results for blocks of sizes 200, 100, 50, 25, and 5 positions.

The position from which the parsimony score on the gene trees becomes lower than that on the species tree is position 221. Interestingly, the 3′ half of the *rps11* gene starts at position 220, indicating that, indeed, this part of the gene had evolved down a tree other than the species tree.

The ML Criterion for Phylogenetic Networks

In the context of optimization criteria for phylogeny reconstruction, we have recently introduced a ML framework for evaluating and reconstructing phylogenetic networks (Jin et al. 2006a). Like the ML criterion for phylogenetic trees, this framework views a phylogenetic network from a probabilistic perspective as a generative model, and the phylogenetic network that maximizes the likelihood of the sequences at its leaves is sought. Furthermore, in a similar manner to that of defining the parsimony of networks, the ML criterion for phylogenetic networks is defined in terms of the trees contained inside the networks (maximizing or summing over all trees). Our preliminary results indicate that the ML framework is a promising approach as well. Yet, the parsimony criterion currently outperforms it, in terms of computational requirements as well as accuracy of the inferred HGT events. However, it is important to note that the accuracy issue is just an artifact of our initial (and naive) way of estimating the parameters associated with the branches of the networks and trees. Once more appropriate stochastic models of evolution down phylogenetic networks are defined and used; we expect the relative performance of the two criteria to be similar to that in the context of phylogenetic trees.

Acknowledgments

This work was supported in part by the Rice Terascale Cluster funded by National Science Foundation under grant EIA-0216467, Intel, and HP. L.N. was supported in part by the Department of Energy grant DE-FG02-06ER25734, the National Science Foundation grant CCF-0622037, the George R. Brown School of Engineering Roy E. Campbell Faculty Development Award, and the Department of Computer Science at Rice University. We are grateful to Associate Editor Dan Graur and the 2 anonymous reviewers for helpful comments and suggestions.

Literature Cited

Addario-Berry L, Hallett MT, Lagergren J. 2003. Towards identifying lateral gene transfer events. In: Altman RB, Dunker AK, Hunter L, Klein TE, editors. Proc 8th Pacific Symp on Biocomputing (PSB03). p. 279–290.

Bergthorsson U, Adams KL, Thomason B, Palmer JD. 2003. Widespread horizontal transfer of mitochondrial genes in flowering plants. Nature. 424:197–201.

Bergthorsson U, Richardson A, Young GJ, Goertzen L, Palmer JD. 2004. Massive horizontal transfer of mitochondrial genes from diverse land plant donors to basal angiosperm Amborella. Proc Natl Acad Sci USA. 101:17747–17752.

Blanchette M, Bourque G, Sankoff D. 1997. Breakpoint phylogenies. In: Miyano S, Takagi T, editors. Genome Informatics. Tokyo: University Academy Press. p. 25–34.

Boc A, Makarenkov V. 2003. New efficient algorithm for detection of horizontal gene transfer events. In: Benson G, Page R, editors. Proc. 3rd Int'l Workshop Algorithms in Bioinformatics (WABI03). Vol. 2812. Springer-Verlag. p. 190–201.

Brown JR. 2003. Ancient horizontal gene transfer. Nat Rev Genet. 4:121–132.

Brown JR, Douady CJ, Italia MJ, Marshall WE, Stanhope MJ. 2001. Universal trees based on large combined protein sequence data sets. Nat Genet. 28:281–285.

Bull JJ, Huelsenbeck JP, Cunningham CW, Swofford D, Waddell P. 1993. Partitioning and combining data in phylogenetic analysis. Syst Biol. 42(3):384–397.

Chippindale PT, Wiens JJ. 1994. Weighting, partitioning, and combining characters in phylogenetic analysis. Syst Biol. 43(2):278–287.

Cunningham CW. 1997. Can three incongruence tests predict when data should be combined? Mol Biol Evol. 14:733–740.

Daubin V, Gouy M, Perriere G. 2002. A phylogenomic approach to bacterial phylogeny: evidence of a core of genes sharing a common history. Genome Res. 12:1080–1090.

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

Daubin V, Ochman H. 2004. Quartet mapping and the extent of lateral transfer in bacterial genomes. Mol Biol Evol. 21(1): 86–89.

Day WHE. 1983. Computationally difficult parsimony problems in phylogenetic systematics. J Theor Biol. 103:429–438.

Delwiche CF, Palmer JD. 1996. Rampant horizontal transfer and duplicaion of rubisco genes in eubacteria and plastids. Mol Biol Evol. 13(6):873–882.

de Queiroz A, Donoghue MJ, Kim J. 1995. Separate versus combined analysis of phylogenetic evidence. Annu Rev Ecol Syst. 25:657–681.

Doolittle WF. 1999a. Lateral genomics. Trends Biochem Sci. 24(12):M5–M8.

Doolittle WF. 1999b. Phylogenetic classification and the universal tree. Science. 284:2124–2129.

Doolittle WF, Boucher Y, Nesbo CL, Douady CJ, Andersson JO, Roger AJ. 2003. How big is the iceberg of which organellar genes in nuclear genomes are but the tip? Philos Trans R Soc Lond B Biol Sci. 358:39–57.

Eisen JA. 2000a. Assessing evolutionary relationships among microbes from whole-genome analysis. Curr Opin Microbiol. 3:475–480.

Eisen JA. 2000b. Horizontal gene transfer among microbial genomes: new insights from complete genome analysis. Curr Opin Genet Dev. 10(6):606–611.

Fitch WM. 1971. Toward defining the course of evolution: minimum change for a specified tree topology. Syst Zool. 20:406–416.

Foulds LR, Graham RL. 1982. The steiner problem in phylogeny is NP-complete. Adv Appl Math. 3:43–49.

Garcia-Vallve S, Guzman E, Montero MA, Romee A. 2003. HGT-DB: a database of putative horizontally transferred genes in prokaryotic complete genomes. Nucleic Acids Res. 31: 187–189. Available from: http://www.fut.es/ ~ debb/HGT/. Accessed on June 10, 2006.

Hallett MT, Lagergren J. 2001. Efficient algorithms for lateral gene transfer problems. In: Lengauer T, editor. Proc 5th Annu Int'l Conf Comput Mol Biol. (RECOMB01). New York: ACM Press. p. 149–156.
Hao W, Golding GB. 2004. Patterns of bacterial gene movement. Mol Biol Evol. 21(7):1294–1307.
Hartigan JA. 1973. Minimum mutation fits to a given tree. Biometrics. 29:53–65.
Hein J. 1990. Reconstructing evolution of sequences subject to recombination using parsimony. Mathematical Biosci. 98: 185–200.
Hein J. 1993. A heuristic method to reconstruct the history of sequences subject to recombination. J Mol Evol. 36:396–405.
Huelsenbeck JP, Bull JJ, Cunningham CW. 1996. Combining data in phylogenetic analysis. Trends Ecol Evol. 11(4):151–157.
Huynen MA, Bork P. 1998. Measuring genome evolution. Proc Natl Acad Sci USA. 95:5849–5856.
Jain R, Rivera MC, Moore JE, Lake JA. 2002. Horizontal gene transfer in microbial genome evolution. Theor Popul Biol. 61(4):489–495.
Jain R, Rivera MC, Moore JE, Lake JA. 2003. Horizontal gene transfer accelerates genome innovation and evolution. Mol Biol Evol. 20(10):1598–1602.
Jin G, Nakhleh L, Snir S, Tuller T. Forthcoming 2006a. Maximum likelihood of phylogenetic networks. Bioinformatics. 22(21):2604–2611.
Jin G, Nakhleh L, Snir S, Tuller T. Forthcoming 2006b. Parsimony of phylogenetic networks: hardness results and efficient algorithms and heuristics. In: Proceedings of the European Conference on Computational Biology (ECCB).
Judd WS, Olmstead RG. 2004. A survey of tricolpate (eudicot) phylogenetic relationships. Am J Bot. 91:1627–1644.
Kunin V, Goldovsky L, Darzentas N, Ouzounis CA. 2005. The net of life: reconstructing the microbial phylogenetic network. Genome Res. 15:954–959.
Kurland CG. 2000. Something for everyone—horizontal gene transfer in evolution. Embo Reports. 1(2):92–95.
Kurland CG, Canback B, Berg OG. 2003. Horizontal gene transfer: a critical view. Proc Natl Acad Sci USA. 100(17): 9658–9662.
Lake JA, Jain R, Rivera MC. 1999. Mix and match in the Tree of Life. Science. 283:2027–2028.
Lawrence JG, Ochman H. 1997. Amelioration of bacterial genomes: rates of change and exchange. J Mol Evol. 44: 383–397.
Lawrence JG, Ochman H. 2002. Reconciling the many faces of lateral gene transfer. Trends in Microbiol. 10(1):1–4.
Lerat E, Daubin V, Moran NA. 2003. From gene trees to organismal phylogeny in prokaryotes: the case of the γ-proteobacteria. PLoS Biol. 1(1):1–9.
Maddison W. 1997. Gene trees in species trees. Syst Biol. 46(3):523–536.
Matte-Tailliez O, Brochier C, Forterre P, Philippe H. 2002. Archaeal phylogeny based on ribosomal proteins. Mol Biol Evol. 19(5):631–639.
McClilland M, Sanderson KE, Clifton SW, et al. (35 co-authors). 2004. Comparison of genome degradation in Paratyphi A and Typhi, human-restricted serovars of Salmonella enterica that cause typhoid. Nat Genet. 36(12):1268–1274.
Medigue C, Rouxel T, Vigier P, Henaut A, Danchin A. 1991. Evidence for horizontal gene transfer in E. coli speciation. J Mol Biol. 222:851–856.
Michelangeli FA, Davis JI, Stevenson D Wm. 2003. Phylogenetic relationships among poaceae and related families as inferred from morphology, inversions in the plastid genome, and sequence data from mitochondrial and plastid genomes. Am J Bot. 90:93–106.
Moret BME, Nakhleh L, Warnow T, Linder CR, Tholse A, Padolina A, Sun J, Timme R. 2004. Phylogenetic networks: modeling, reconstructibility, and accuracy. IEEE/ACM Trans Comput Biol Bioinform. 1(1):13–23.
Mower JP, Stefanovic S, Young GJ, Palmer JD. 2004. Gene transfer from parasitic to host plants. Nature. 432:165–166.
Nakamura Y, Itoh T, Matsuda H, Gojobori T. 2004. Biased biological functions of horizontally transferred genes in prokaryotic genomes. Nat Genet. 36(7):760–766.
Nakhleh L, Jin G, Zhao F, Mellor-Crummey J. 2005. Reconstructing phylogenetic networks using maximum parsimony. In: Markstein V, editor. Proceedings of the 2005 IEEE Computational Systems Bioinformatics Conference (CSB2005); August. p. 93–102.
Nakhleh L, Ruths D, Wang LS. 2005. RIATA-HGT: a fast and accurate heuristic for reconstrucing horizontal gene transfer. In: Wang L, editor. Proceedings of the Eleventh International Computing and Combinatorics Conference (COCOON 05). Springer-Verlag, Berlin, Heidelberg. Lecture Notes in Computer Science; Vol. 3595. p. 84–93.
Ochman H, Lawrence JG, Groisman EA. 2000. Lateral gene transfer and the nature of bacterial innovation. Nature. 405(6784): 299–304.
Olmstead RG, Sweere JA. 1994. Combining data in phylogenetic systematics: an empirical approach using three molecular data sets in the Solanaceae. Syst Biol. 43(4):467–481.
Sankoff D, Blanchette M. 1998. Multiple genome rearrangement and breakpoint phylogeny. J Comput Biol. 5:555–570.
Shimodaira H, Hasegawa M. 1999. Multiple comparisons of log-likelihoods with applications to phylogenetic inference. Mol Biol Evol. 16:1114–1116.
Teichmann SA, Mitchison G. 1999. Is there a phylogenetic signal in prokaryote proteins? J Mol Evol. 49:98–107.
Welch RA, Burland V, Plunkett G, et al. (19 co-authors). 2002. Extensive mosaic structure revealed by the complete genome sequence of uropathogenic Escherichia coli. Proc Natl Acad Sci USA. 99:17020–17024.
Wiens JJ. 1998. Combining data sets with different phylogenetic histories. Syst Biol. 47:568–581.

Dan Graur, Associate Editor

Accepted October 20, 2006