BIOINFORMATICS

Vol. 23 ECCB 2006, pages e123–e128
doi:10.1093/bioinformatics/btl313

Efficient parsimony-based methods for phylogenetic network reconstruction

Guohua Jin[1], Luay Nakhleh[1], Sagi Snir[2,*] and Tamir Tuller[3,†]

[1]Department of Computer Science, Rice University, Houston, TX, USA, [2]Department of Mathematics, University of California, Berkeley, CA, USA and [3]School of Computer Science, Tel Aviv University, Tel Aviv, Israel

ABSTRACT

Motivation: Phylogenies—the evolutionary histories of groups of organisms—play a major role in representing relationships among biological entities. Although many biological processes can be effectively modeled as tree-like relationships, others, such as hybrid speciation and horizontal gene transfer (HGT), result in networks, rather than trees, of relationships. Hybrid speciation is a significant evolutionary mechanism in plants, fish and other groups of species. HGT plays a major role in bacterial genome diversification and is a significant mechanism by which bacteria develop resistance to antibiotics. Maximum parsimony is one of the most commonly used criteria for phylogenetic tree inference. Roughly speaking, inference based on this criterion seeks the tree that minimizes the amount of evolution. In 1990, Jotun Hein proposed using this criterion for inferring the evolution of sequences subject to recombination. Preliminary results on small synthetic datasets. Nakhleh *et al.* (2005) demonstrated the criterion's application to phylogenetic network reconstruction in general and HGT detection in particular. However, the naive algorithms used by the authors are inapplicable to large datasets due to their demanding computational requirements. Further, no rigorous theoretical analysis of computing the criterion was given, nor was it tested on biological data.

Results: In the present work we prove that the problem of scoring the parsimony of a phylogenetic network is NP-hard and provide an improved fixed parameter tractable algorithm for it. Further, we devise efficient heuristics for parsimony-based reconstruction of phylogenetic networks. We test our methods on both synthetic and biological data (rbcL gene in bacteria) and obtain very promising results.

Contact: ssagi@math.berkeley.edu

1 INTRODUCTION

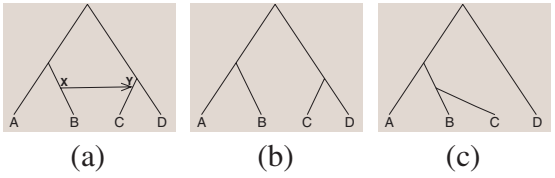
Phylogenetic networks are a special class of directed acyclic graphs that models evolutionary histories when trees are inappropriate, such as in the cases of horizontal gene transfer (HGT) and hybrid speciation (Linder *et al*., 2004; Moret *et al*., 2004; Makarenkov *et al*., 2006). Figure 1a shows a phylogenetic network on four species with a single HGT event. In HGT, genetic material is transferred from one lineage to another, as in Figure 1a. In an evolutionary scenario involving horizontal transfer, certain sites (specified by a specific substring within the DNA sequence of the species into which the horizontally transferred DNA was inserted) are inherited through horizontal transfer from another species (as in Fig. 1c), while all others are inherited from the parent (as in Fig. 1b). Thus, each site evolves down one of the trees induced by (or contained in) the network. Similar scenarios arise in the cases of other reticulate evolution events (such as hybrid speciation and interspecific recombination). Hybrid speciation is a significant evolutionary mechanism in plants, fish and other groups of species (Linder and Rieseberg, 2004). HGT plays a major role in bacterial genome diversification (Doolittle *et al*., 2003; Eisen, 2000) and is a significant mechanism by which bacteria develop resistance to antibiotics (Paulsen *et al*., 2003). To facilitate evolutionary analysis of these groups of organisms, developing accurate criteria for reconstructing phylogenetic networks and efficient algorithms for inference based on these criteria are imperative. A large number of publications have been introduced in recent years about various aspects of phylogenetic networks; see (Linder *et al*., 2004; Makarenkov *et al*., 2006) for detailed surveys.

Maximum parsimony (MP) is one of the most widely used criteria for phylogenetic tree analysis. It is based on a minimum-evolution principle, compares well with other accurate criteria and has a host of efficient algorithms for solving problems based on it (Fitch, 1971; Gusfield, 1991). In 1990, Hein observed that the criterion could be extended to detect recombination (Hein, 1990, 1993). He observed that each individual site in a set of sequences labeling a network evolves down a tree contained in the network (e.g. the trees in Fig. 1b and 1c are contained in the network shown in Fig. 1a). Following this observation, Nakhleh *et al*. (2005) formulated the parsimony criterion for inferring and evaluating phylogenetic networks. The HGT reconstruction problem seeks an optimal set of edges whose addition to a given species tree results in an optimal network that explains the given gene data. In the context of parsimony, we refer to this problem as the fixed-tree MP phylogenetic network problem, or FTMPPN. Solving this problem entails scoring the parsimony of a phylogenetic network leaf-labeled by a set of sequences; we refer to this problem as the parsimony score of phylogenetic network problem or PSPN. Nakhleh *et al*. (2005) used a straightforward algorithm (exponential in the number of reticulation edges) for solving the PSPN problem and exhaustively searched all trees for solving the FTMPPN problem. Further, they left open the question of the computational complexity of these problems.

In the present study, we prove that the PSPN problem is NP-hard. However, on the positive side, we give an efficient algorithm for the problem and bound its running time to prove that it is fixed parameter tractable (Downey and Fellows, 1995). The algorithm has very good performance in practice, as we show, and was integrated as part of efficient heuristics for the FTMPPN problem. Further, we devise new heuristics for the FTMPPN problem and show through experiments on biological as well as synthetic data that

*To whom correspondence should be addressed.

†The authors wish it to be known thats in their opinion, all the authors should be regarded as Joint First Authors.



(a) (b) (c)

Fig. 1. (**a**) A phylogenetic network with a single HGT even from *X* to *Y*. (**b**) The underlying organismal (species) tree. (**c**) The tree of a horizontally transferred gene.

the heuristics are efficient in practice, while maintaining a high accuracy. The biological dataset we analyze includes the rbcL gene in plastids, cyanobacteria and proteobacteria. A set of HGTs was hypothesized for this dataset (Delwiche and Palmer, 1996).

A large body of work has been introduced in recent years to address phylogenetic network reconstruction and evaluation. In general, three categories of non-treelike models have been addressed, all of which have been introduced under the umbrella concept of phylogenetic networks. However, major differences exist among the three categories. Splits networks are graphical models that capture incompatibilities in the data due to various factors, not necessarily HGT or hybrid speciation. Phylogenetic networks are the extension of phylogenetic trees to enable the modeling of reticulation events, such as HGT and hybrid speciation [these are also called reticulate networks in (Huson and Bryant, 2006)]. The third category is that of recombination networks, which are used to model the evolution of haplotypes and genes at the population level. See Linder *et al*. (2004) and Makarenkov *et al*. (2006) for detailed surveys of the various phylogenetic network models and methodologies. Phylogenetic networks that we address in this work belong to the second category.

2 PARSIMONY OF NETWORKS

2.1 Preliminaries and definitions

Let $T = (V, E)$ be a tree, where V and E are the tree nodes and tree edges, respectively, and let $L(T)$ denote its leaf set. Further, let X be a set of taxa (species). Then, T is a phylogenetic tree over X if there is a bijection between X and $L(T)$. Henceforth, we will identify the taxa set with the leaves they are mapped to, and let $[n] = \{1, \ldots, n\}$ denote the set of leaf-labels. A tree T is said to be rooted if the set of edges E is directed and there is a single distinguished internal vertex r with in-degree 0. We denote by T_v the subtree rooted at v induced by the tree edges. A function $\lambda : [n] \rightarrow \{0, 1, \ldots, \Sigma - 1\}$ is called a state assignment function over the alphabet Σ for T. We say that function $\hat{\boldsymbol{\lambda}} : V(T) \rightarrow \{0, 1, \ldots, \Sigma - 1\}$ is an extension of λ on T if it agrees with λ on the leaves of T. In a similar way, we define a function $\lambda^k : [n] \mapsto \{0, 1, \ldots, \Sigma - 1\}^k$ and an extension $\hat{\boldsymbol{\lambda}}^k : V(T) \mapsto \{0, 1, \ldots, \Sigma - 1\}^k$. The latter function is called a labeling of T. We write $\hat{\boldsymbol{\lambda}}^k(v) = s$ to denote that sequence s is the label of the vertex v. Every position $1 \leq i \leq k$ denotes a site in the sequence. Given a labeling $\hat{\boldsymbol{\lambda}}^k$, let $d_e(\hat{\boldsymbol{\lambda}}^k)$ denote the Hamming distance between the two sequences labeling the two endpoints of the edge $e \in E(T)$.

A phylogenetic network $N = N(T) = (V', E')$ over the taxa set X is derived from $T = (V, E)$ by adding a set H of edges to T, where each edge $h \in H$ is added as follows: (1) split an edge $e \in E$ by adding new node, v_e; (2) split an edge $e' \in E$ by adding new node, $v_{e'}$ and (3) finally, add a directed reticulation edge from v_e to $v_{e'}$. Phylogenetic networks must satisfy additional temporal constraints (Moret *et al*., 2004). Finally, we denote by $T(N)$ the set of all trees contained inside network N. Each such tree is obtained by the following two steps: (1) for each node of in-degree 2, remove one of the incoming edges, and then (2) for every node x of in-degree and out-degree 1, whose parent is u and child is v, remove node x and its two adjacent edges, and add a new edge from u to v. Figure 1 shows a network and the two trees it contains. For a network N and a node $v \in V(N)$, N_v denotes the graph induced by the nodes reachable from v.

2.2 Parsimony of phylogenetic networks

We begin by reviewing the parsimony criterion for phylogenetic trees. Given a phylogenetic tree with a labeling (sequences) of its leaves, the idea is to add labels to its internal nodes such that the sum of Hamming distances along all the edges of the tree is minimized. More formally,

PROBLEM 1. *Parsimony Score of Phylogenetic Trees (PSPT)*

Input: A 3-tuple (S, T, λ^k), *where T is a phylogenetic tree and* λ^k *is the labeling of L(T) by the sequences in S.*

Output: *The extension* $\hat{\boldsymbol{\lambda}}^k$ *that minimizes* $\sum_{e \in E(T)} d_e(\hat{\boldsymbol{\lambda}}^k)$.

We define the parsimony score for (S, T, λ^k), $\text{pars}(S, T, \lambda^k)$, as the value of the sum of Hamming distances along the tree's edges and $\text{pars}(S, T, \lambda^k, i)$ as this sum of Hamming distances for site i only. In other words, $\text{pars}(S, T, \lambda^k) = \sum_{1 \leq i \leq k} \text{pars}(S, T, \lambda^k, i)$. Problem 1 has a polynomial time dynamic programming type algorithm originally devised for binary characters and binary trees by Fitch (1971) and later extended to arbitrary degree trees and multi-state characters by Sankoff (1975). The algorithm finds an optimal assignment (i.e. $\hat{\boldsymbol{\lambda}}^k$) for each site separately.

Since Fitch's algorithm is a basic building block in this paper, we hereby describe it. As mentioned above, the input to the problem is a tree T and a single character $C = \lambda^1$. The algorithm finds $pars(\{1, 0\}, T, C)$, the optimal assignment to internal nodes of T, in two phases: (1) assigning values to internal nodes in a bottom-up fashion and (2) eliminating the values determined in the previous phase in a top-down fashion. Specifically, phase (1) proceeds as follows: for a node v with children v_1 and v_2 whose values $A(v_1)$ $A(v_2)$ have been determined[1],

$$A(v) = \begin{cases} A(v_1) \cap A(v_2) & \text{if } A(v_1) \cap A(v_2) \neq \varnothing \\ A(v_1) \cup A(v_2) & \text{otherwise.} \end{cases}$$

Phase (2) proceeds as follows: for a node v whose parent $f(v)$ has already been processed:

$$B(v) = \begin{cases} \sigma \in A(v) \cap A(f(v)) & \text{if } A(v) \cap A(f(v)) \neq \phi \\ \sigma \in A(v) & \text{otherwise} \end{cases}.$$

The algorithm applies to binary trees and extends in a straightforward manner to arbitrary k-degree trees by a slight modification to phase (2): at each node v, $B(v)$ is a state that is a member of a majority of all $A(v_i)$ for all children i and the ancestor of v. The following lemma will be used later.

[1] $A(v)$ is the set of all labels whose assignment to node v yield the minimum parsimony score of subtree T_v. For leaf node v whose label is x, we have $A(v) = \{x\}$.

LEMMA 1. *Let T be a tree and C a single character over the alphabet* Σ. *Let x be the number of internal nodes v s.t.* $|A(v)| > 1$ *by applying Fitch's algorithm on (T,C). Then* $x < 2 \cdot S^*$, *where* S^* *is the parsimony score of T over C.*

As explained in Section 1 and illustrated in Figure 1, when an HGT event occurs, the evolutionary history of the complete genomes of the organisms is modeled by a phylogenetic network. Nevertheless, the evolutionary history of every site in these genomes is modeled by one of the phylogenetic trees inside the network. This gives rise to the following definition of the parsimony score of phylogenetic networks [as was introduced by Hein (1990, 1993) and formalized by Nakhleh *et al*. (2005)].

DEFINITION 1. *Parsimony Score of Phylogenetic Networks (PSPN)*

Input: *A* 3*-tuple* (S, N, λ^k), *where N is a phylogenetic network and* λ^k *is the labeling of L(N) by the sequences in S.*

Output: *The extension* $\hat{\lambda}^k$ *that minimizes the expression* $\sum_{1 \le i \le k} [min_{T \in T(N)} pars(S, T, \lambda^k, i)]$.

The parsimony score of a network is the value of the sum above. In the next section, we prove that the PSPN problem is NP-hard. Notice that based on Definition 1 the parsimony of each site is computed independently of the other sites, and hence we focus on the case of a single site.

3 THE PSPN PROBLEM

3.1 Hardness of the problem

In the same spirit of MP heuristics for phylogenetic trees, a crucial part of heuristics for solving the MP problem on phylogenetic networks involves solving the PSPN problem. The decision version of the problem for the case of a single binary site is defined as follows.

PROBLEM 2. *(PSPN*1*)*

Input: *A phylogenetic network* $N = N(T) = (V', E')$ *with binary labeling of length* 1, *and an integer k.*

Question: *Is the MP score of the network* $\le k$?

We prove the hardness of the PSPN1 problem by a reduction from the Maximum 2-Satisfiability (max–2–sat) problem (Garey and Johnson, 1979). In max–2–sat the input is a set of clauses, each with two literals. The goal is to find an assignment that satisfies a maximum number of these clauses.

Due to space limitations, we only give a very general outline of our proof. Let 'True-True' denote a clause that has no negated literals, 'True-False' denote a clause that has exactly one negated literal and 'False-False' denote a clause in which both literals are negated. For each of these three types of clauses, we generate subnetworks whose optimal parsimony score is 3, and such that this score is determined by the labeling of two nodes (roots) in each such subnetwork. These nodes correspond to the literals in the max–2–sat problem. Each such node (literal) should be connected to all the subnetworks (clauses) in which it appears in the max–2–sat problem. Using this reduction we prove the following theorem:

THEOREM 1. *The PSPN*1 *problem is NP-hard.*

Since *Max–2–sat* is hard even for inputs where each variable is restricted to appear at most 12 times, the *PSPN*1 problem is NP-hard even for networks of bounded degrees (where each node has at most 12 children).

3.2 An improved FPT algorithm

DEFINITION 2. *A reticulation edge* $(u \to v)$ *is called a lowest reticulation edge (or just a lowest edge) if there is no reticulation edge incident with any node in either* T_u *or* T_v.

LEMMA 2. *For every phylogenetic network, there exists a lowest edge.*

This lemma follows from the fact that phylogenetic networks are acyclic and satisfy additional temporal constraints (Moret *et al*., 2004).

COROLLARY 1. *Let* $(u \to v)$ *be a lowest edge. Then both* N_u *and* N_v *are trees.*

The algorithm Net2Trees of Nakhleh *et al*. (2005) enumerates all the 2^B possible trees contained inside a given network with B reticulation edges and calculates the parsimony score of each tree by running Fitch's algorithm (Fitch, 1971) in $O(n|\Sigma|)$ time. The optimal score among all trees contained inside the network is then returned. The total running time is $2^B \cdot O(n|\Sigma|)$, which, for a fixed B is polynomial. However, a very simple example demonstrates that this running time can be unnecessarily extremely high. Consider a site with a single observed state. Obviously, the underlying tree yields the optimal assignment with score 0. In contrast the naïve algorithm of Nakhleh *et al*. (2005) will run in time exponential in B (and in n in a worst-case scenario).

We now present our improvement to the algorithm from Nakhleh *et al*. (2005) for computing the optimal score of a network N. By Lemma 2, there exists a lowest edge $e = (u \to v)$ in N and by Corollary 1 the subnetworks reachable from both endpoints u and v are trees. Therefore we can compute *A(u)* and *A(v)* by Fitch's algorithm. The following lemma is fundamental for the algorithm correctness.

LEMMA 3. *Let* $e = (u \to v)$ *be a lowest edge in a network N for which A(u) and A(v) have already been computed. Also assume the resulting tree contains e. Then*,

(1) *If* $A(u) \subseteq A(v)$, *there will be no mutation on e.*

(2) *If* $A(u) \cap A(v) = \phi$, *there will be a mutation on e.*

In the cases that are not covered by Lemma 3, we say that v is *uncertain*. Lemma 3 gives rise to the recursive algorithm, PSPN(N), for computing the optimal score of N, as outlined in Figure 2.

The correctness of the algorithm is implied by the construction and Lemma 3. A reticulation edge $(u \to v)$ is automatically taken into the tree only if it yields no mutation and is automatically rejected from the tree if it necessarily leads to a mutation. In all other cases v is uncertain, and both cases are considered.

The algorithm recurs only on reticulation edges $(u \to v)$ where v is uncertain. Given that $|A(v)| > 1$, and by Lemma 1, it follows that the number of such nodes is at most twice the optimal score of N.

THEOREM 2. *The running time of the improved FPT algorithm is* $O(n \cdot 2^{\text{OPT}(N)})$, *where n is the number of nodes in the network and opt(N) is the optimal parsimony score of the site (under consideration) on the network.*

PSPN(N=(V',E'))

1. If N is not a tree
   a. Find a lowest reticulation edge $e = (u \rightarrow v)$ in N;
   b. Let e' be the edge between v and its ancestral node on the tree edge;
   c. By Fitch's algorithm, compute the optimal assignment A of u and v;
   d. If $A(u) \cap A(v) = \emptyset$ then $opt = PSPN(V', E' \setminus e)$;
   e. else if $A(u) \subseteq A(v)$ then $opt = PSPN(V', E' \setminus e')$;
   f. else
      (1) $opt = PSPN(V', E' \setminus e)$;
      (2) $opt' = PSPN(V', E' \setminus e')$;
      (3) if $opt' < opt$ then $opt \leftarrow opt'$;
   g. return opt;
2. else
   a. Let T be the resulting tree;
   b. return $Fitch(T)$;

Fig. 2. The improved FPT algorithm for the PSPN problem.

4 THE FTMPPN PROBLEM

Finally, we consider the fixed-tree MP on phylogenetic networks (FTMPPN) problem (Nakhleh *et al*., 2005). In this problem, given an organismal (species) tree, the objective is to compute an additional set of edges whose addition to the tree yields a phylogenetic network that explains the horizontal gene transfer events which occurred during the evolutionary history of the sequences. This problem arises in situations when the underlying organismal tree is known. For example, Lerat *et al*. (2003) reported a well-supported organismal phylogeny reconstructed from about 100 'core' genes in γ-Proteobacteria. Completing this phylogenetic tree into a network based on the whole genomes of these organisms amounts to detecting HGT events that occurred in the γ-Proteobacteria group.

Since the actual number of the HGT events as well as their locations are not known, parsimony is used as the optimality criterion for the search. Nakhleh *et al*. (2005) showed that solving this problem accurately detects the HGT events in a sequence dataset. However, since their goal was to study the quality of the approach rather than the efficiency of computing it, they had a brute-force implementation that took almost 10 h on datasets with only two HGT events. Since this is infeasible in practice, we devise simple, yet efficient and accurate, heuristics for solving the FTMPPN problem and demonstrate, through simulations, its excellent accuracy.

The preliminary results in Nakhleh *et al*. (2005) showed that the optimal phylogenetic networks with k reticulation edges could always be obtained from the optimal phylogenetic networks with $k-1$ reticulation edges. Based on this observation, we implemented a branch and bound heuristic (B&B) in which at each step of the search only 'best' networks are retained. Further, to find the optimal phylogenetic networks with k reticulation edges, we conducted search based only on the optimal ones with $k-$reticulation edges. This cuts the time significantly, while maintaining excellent accuracy (in terms of the optimality of the score computed by the heuristic compared with that of the model network), as we will show.

To gain further improvements in time, we extended the B&B heuristic by inspecting Hamming distances on the tree edges; we call this heuristic B&B (Hamming). This heuristic divides the sequences (that label the nodes of the species tree) into blocks. Then, the heuristic applies Fitch's algorithm and labels the internal nodes of the tree. Next, for each edge, it computes the Hamming distance for each of the blocks and normalizes it by the average Hamming distance over all blocks along the same edge. Finally, for each edge we compute the difference between the maximum and minimum values of normalized Hamming distances over all blocks and use this value as a criterion for finding candidate edges. Finally, the search for tree edges among which to add HGT events is done in the (reduced) space of candidate edges. The rationale behind this approach is that for DNA segments that were horizontally transferred (rather than inherited down the species tree) the parsimony score on the species tree should be higher than that of segments that evolved down the species tree. The reason for this is that the species tree does not model the evolution of horizontally transferred DNA segments, and hence that tree should not be a 'good' model for these segments (which translates into high parsimony scores).

5 EMPIRICAL PERFORMANCE

5.1 Data and methods

For the biological data, we considered a 15-taxon dataset of plastids, cyanobacteria and proteobacteria, which is a subset of the dataset considered by Delwiche and Palmer (1996) and for which multiple HGT events were conjectured by the authors. The 15-taxon rbcL dataset consists of two sequences from the proteobacteria group, two from cyanobacteria, one from green plastids, one from red plastids, one cyanophora and four Form II rubisco sequences. For this dataset, we obtained the species tree that was reported in Delwiche and Palmer (1996) and analyzed the rubisco gene rbcL of these 15 organisms. The gene dataset consists of 15 aligned amino acid sequences, each of length 532 (the alignment is available at http://www.life.umd.edu/labs/delwiche/alignments/rbcLgb7-95.distrib.txt). We note that our method finds both the HGT edges and the positions that are involved in each HGT.

For the synthetic data, we used the following protocol to generate them. We used the r8s tool Sanderson (r8s) to generate a random birth-death phylogenetic tree on 20 taxa. The r8s tool generates molecular clock trees; we deviated the tree from this hypothesis by multiplying each edge in the tree by a number randomly drawn from an exponential distribution. The expected evolutionary diameter (longest path between any two leaves in the tree) is 0.2. Such diameter can simulate the evolution of a set of plants [see Bergthorsson (2004) for example of horizontal transfers in plants].

We then generated five model phylogenetic networks by adding 1, 2, 3, 4 and 5 reticulation edges (simulating HGT events) to the model tree. For each of the five phylogenetic networks, we used the Seq-gen tool Rambaut and Grassly (1997) to evolve 26 datasets of DNA sequences of length 1500 down the organismal tree and DNA sequences of length 500 down the other tree contained inside the network (the one that exhibits all HGT events). Both sequence datasets were evolved under the K2P+γ model of evolution, with shape parameter 1 Kimura (1980). Finally, we concatenated the two datasets.

To analyze the data, we have implemented the B&B as well as the B&B (Hamming) heuristics for solving the FTMPPN problem. As

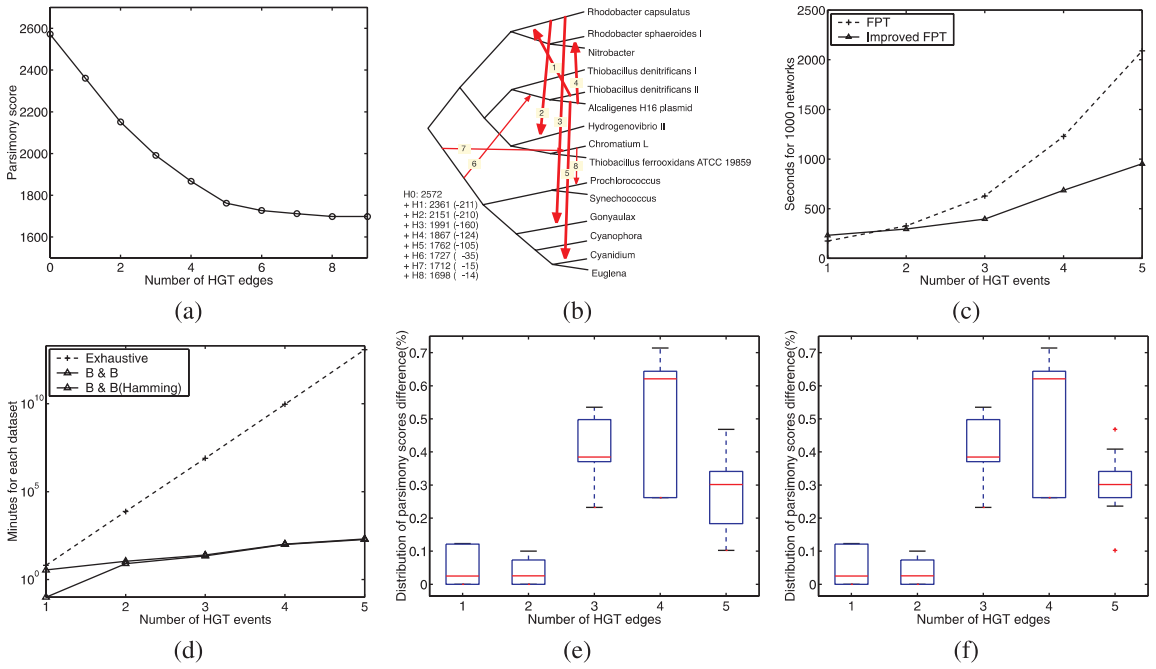


Fig. 3. (a) and (b) show results on the biological dataset, whereas (c)–(f) show results on the synthetic datasets. (**a**) The improvement in the parsimony score as more HGT edges are added to the 15-taxon organismal tree. (**b**) The phylogenetic network obtained by adding the HGT edges that led to the best improvement in the parsimony score. (**c**) The actual computational time (averaged over all 26 runs) of the naïve FPT algorithm of (Nakhleh *et al*., 2005) and our improved FPT algorithm for solving the PSPN problem, as a function of the number of HGT edges in the network. (**d**) actual computational time (averaged over all 26 runs) taken by the heuristics and the exhaustive search method (on a log scale); these times were taken to solve the FTMPPN problem using the three methods. (**e**) and (**f**): percent difference in parsimony scores (for all runs) between optimal networks computed by the heuristics and the model networks, shown with whisker-and-box plots.

these two heuristics entail scoring the parsimony of a phylogenetic network, we have implemented the naive algorithm, introduced in Nakhleh *et al*. (2005) and referred to as 'FPT' in the results section, as well as the new improved one, described in Section 3.2 and referred to as 'Improved FPT' in the Results section, and compared their performance in terms of time. In our analysis, we aimed to investigate two main questions: (1) How do the new heuristics for solving the FTMPPN problem perform with respect to identifying the correct number of HGT events as well as their actual locations on the organismal trees? (2) How does the Improved FPT algorithm perform, in terms of actual running time, compare with the FPT algorithm? In both the biological as well as synthetic data, the organismal trees were known. For the biological data, we compared our results against the HGT events conjectured by the authors, and for the simulated data, we compared our results against the (known) correct solutions.

5.2 Results and analysis

5.2.1 Biological data Figure 3a shows the parsimony scores of the most parsimonious networks with different number of horizontal gene transfer edges. We computed the weighted parsimony scores, using five different amino acid substitution matrices: PAM120, PAM250, BLOSUM45, BLOSUM62 and IDENTITY. The results based on these five matrices were almost identical, and due to space constraints we show only the results obtained using the IDENTITY matrix. Figure 3a shows clearly that parsimony scores drop dramatically when the first 5 or 6 potential HGT edges are added to the species tree. The decrease then becomes insignificant, and no decrease at all is achieved after adding the eighth edge. The edges that resulted in the optimal decrease are shown by directed edges posited on the species tree in Figure 3b, with each of the edges representing a potential transfer of the rbcL gene (the numbers associated with the directed edges represent the order in which they were added). Figure 3b also listed the parsimony score of the most parsimonious network after adding each of the HGT edges. Row '$+Hi$' corresponds to the network after adding *ith* HGT edge into the existing network, while '*H*0' represents the original species tree. The score changes from the previous networks are given inside the parentheses. It is clear that among the 15 taxa, the first five HGT edges are significant, while the others do not result in a significant improvement in the parsimony score, if any at all. The first three HGT edges group the Form II Rubisco together and separates them from the rest (Form I Rubisco). The other two HGT edges are placed between *Cyanidium*, a Red Plastid, and one of the proteobacteria, and between *Alcaligene H16 plasmid* and *Rhodobacter sphaeroides I*. These two HGT edges place the two proteobacteria close to the red plastid. These five edges were conjectured in (Delwiche and Palmer, 1996).

5.2.2 Synthetic data In the first set of experiments, we compared the performances (in terms of actual computational time) of the naive FPT algorithm (Nakhleh *et al*., 2005) and our new improved FPT algorithm. The results are summarized in Figure 3c. The results

show that except for the case of a single HGT event, the improved FPT algorithm becomes much faster than the naive one as the number of HGT edges increases. In particular, for the case of 5 HGT events, the improvement is larger than a factor of 2. More importantly, as the number of HGT events increases, the improvement becomes much more significant (indicated by the widening gap between the two curves in Figure 3c). In the second set of experiments, we studied the performance of the two heuristics B&B and B&B (Hamming) for solving the FTMPPN problem. We compared the time taken by these heuristics with the time the exhaustive search of Nakhleh *et al*. (2005) would take; we had to estimate this latter time, since it would take probably years to perform an exhaustive search on all networks with more than two HGT events. Further, we compared the parsimony scores of the optimal networks computed by these heuristics, PSI, with the parsimony score of the model network, PSM, by the formula (PSM − PSI)/PSM %. This is the value referred to as parsimony score difference(%) in Figure 3e and f. Figure 3d shows drastic improvements in the time achieved by the two heuristics. The exact times in minutes using the B&B heuristic for the networks with 1, 2, 3, 4 and 5 HGT events are 3.5, 11, 25, 103 and 206, respectively. The exact times in minutes using the B&B (Hamming) heuristic for the networks with 1, 2, 3, 4 and 5 HGT events are 0.1, 8, 22, 100 and 192, respectively. On the other hand, the estimated time in minutes using an exhaustive search in the network space are 6.5, 7.3×10^3, 8.0×10^6, 9.5×10^9, and 12.0×10^{12}, respectively. Equally important, the improvement was achieved while maintaining high accuracy in the parsimony scores computed, which is reflected in the negligible score differences plotted in Figures 3e and f. The Figures show that the parsimony scores of the networks inferred by the heuristics fall within 0.7% of the parsimony scores of the model networks. This is a very high accuracy.

6 CONCLUSIONS AND FUTURE WORK

We addressed the parsimony criterion for phylogenetic networks. We proved that PSPN is NP-hard and devised an efficient algorithm for solving it. We also designed efficient heuristics for the FTMPPN problem and tested our methods on biological as well as synthetic data. Our results are very promising and provide a significant contribution toward putting the methodologies for reconstructing and evaluating phylogenetic networks on a par with those for phylogenetic trees.

Though we have assumed site independence, a more realistic model incorporates correlation among neighboring sites. For future work, we will investigate the MP criterion under such models. Since phylogenetic trees are a special case of phylogenetic networks, we expect parsimony's shortcomings on trees (such as the long branch attraction problem) to extend to phylogenetic networks. We will investigate these cases and work on establishing relationships between the MP criterion on the one hand and other network reconstruction criteria on the other. We also intend to analyze others groups of prokaryotic organisms, establish the complexity of the FTMPPN problem and design efficient solutions for it.

ACKNOWLEDGEMENTS

This work was supported in part by the Rice Terascale Cluster funded by NSF under grant EIA-0216467, Intel, and HP. S. S. was supported in part by NSF grant CCR-0105533.

Conflict of Interest: none declared.

REFERENCES

Bergthorsson,U. *et al*. (2004) Massive horizontal transfer of mitochondrial genes from diverse land plant donors to the basal angiosperm Amborella. *Proc. Natl Acad. Sci. USA*, **101**, 17747–17752.

Delwiche,C.F. and Palmer,J.D. (1996) Rampant horizontal transfer and duplicaion of rubisco genes in eubacteria and plastids. *Mol. Biol. Evol*., **13**, 873–882.

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

Downey,R.G. and Fellows,M.R. (1995) Fixed parameter tractability and completeness I: basic theory. *SIAM J. Comput*., **24**, 873–921.

Eisen,J.A. (2000) Assessing evolutionary relationships among microbes from whole-genome analysis. *Curr. Opin. Microbiol*., **3**, 475–480.

Fitch,W. (1971) Toward defining the course of evolution: minimum change for a specified tree topology. *Syst. Zool*., **20**, 406–416.

Garey,M.R. and Johnson,D.S. (1979) *Computer and Intractability*. Bell Telephone Laboratories.

Gusfield,D. (1991) Efficient algorithms for inferring evolutionary history. *Networks*, **21**, 19–28.

Hein,J. (1990) Reconstructing evolution of sequences subject to recombination using parsimony. *Math. 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.

Hochbaum,D.S. (1997) *Approximation Algorithms for NP-Hard Problems*. PWS Publishing Company.

Huson,D.H. and Bryant,D. (2006) Application of phylogenetic networks in evolutionary studies. *Mol. Biol. Evol*. **23**, 254–267.

Kimura,M. (1980) A simple method for estimating evolutionary rates of base substitutions through comparative studies of nucleotide sequences. *J. Mol. Evol*., **16**, 111–120.

Lerat,E. *et al*. (2003) From gene trees to organismal phylogeny in prokaryotes: the case of the γ-proteobacteria. *PLoS Biol*., **1**, 1–9.

Linder,C.R. and Rieseberg,L.H. (2004) Reconstructing patterns of reticulate evolution in plants. *Am. J. Bot*., **91**, 1700–1708.

Linder,C.R. *et al*. (2004) Network (reticulate) evolution: biology, models, and algorithms. In *The Ninth Pacific Symposium on Biocomputing (PSB)*.

Makarenkov,V. *et al*. (2006) Phylogenetic network reconstruction approaches. *Applied Mycology and Biotechnology (Genes, Genomics and Bioinformatics)*, **6**, (in press).

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

Nakhleh,L. *et al*. (2005) Reconstructing phylogenetic networks using maximum parsimony. *Proc. IEEE Comput. Syst. Bioinform. Conf*., **393**, 440–442.

Paulsen,I.T. *et al*. (2003) Role of mobile DNA in the evolution of Vacomycin-resistant Enterococcus faecalis. *Science*, **299**, 2071–2074.

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

Sanderson,M. r8s software package. http://loco.ucdavis.edu/r8s/r8s.html.

Sankoff,D. (1975) Minimal mutation trees of sequences. *SIAM J. Appl. Math*., **28**, 35–42.