

Maximum Likelihood of Phylogenetic Networks

Guohua Jin^a Luay Nakhleh^a Sagi Snir^b Tamir Tuller^{c*}

^a Dept. of Computer Science Rice University Houston, TX, USA ^b Dept. of Mathematics University of California Berkeley, CA, USA ^c School of Computer Science, Tel-Aviv University

Associate Editor: Keith A Crandall

ABSTRACT

Motivation: Horizontal gene transfer (HGT) is believed to be ubiquitous among bacteria, and plays a major role in their genome diversification as well as their ability to develop resistance to antibiotics. In light of its evolutionary significance and implications in human health, developing accurate and efficient methods for detecting and reconstructing HGT is imperative.

Results: In this paper we provide a new HGT-oriented likelihood framework for many problems that involve phylogeny-based HGT detection and reconstruction. Beside the formulation of various likelihood criteria, we show that most of these problems are NP-hard, and offer heuristics for efficient and accurate reconstruction of HGT under these criteria. We implemented our heuristics and used them to analyze biological as well as synthetic data. In both cases, our criteria and heuristics exhibited very good performance with respect to identifying the correct number of HGT events as well as inferring their correct location on the species tree.

Availability: Implementation of the criteria as well as heuristics and hardness proofs are available from the authors upon request. Hardness proofs can also be downloaded at <http://www.cs.tau.ac.il/~tamirtul/MLNET/Supp-ML.pdf>

Contact: Tamir Tuller (tamirtul@post.tau.ac.il)

1 INTRODUCTION

Unlike eukaryotes, which evolve largely through vertical lineal descent, bacteria acquire genetic material through the transfer of DNA segments across species boundaries—a process known as *horizontal gene transfer* (HGT). This process plays a major role in bacterial genome diversification (Doolittle *et al.*, 2003), and is a significant mechanism by which bacteria develop resistance to antibiotics (Paulsen, 2003). In the presence of HGT, the evolutionary history of a set of organisms is modeled by a *phylogenetic network*, which is a directed acyclic graph obtained by positing a set of edges between pairs of the branches of an organismal tree to model the horizontal transfer of genetic material (Moret *et al.*, 2004).

Therefore, to reconstruct an accurate Tree (or Network) of Life and to unravel bacterial genomic complexities, developing accurate criteria and efficient methods for reconstructing

and assessing the quality of phylogenetic networks is imperative. A large body of work has been introduced in recent years to address phylogenetic network reconstruction and evaluation. In general, three categories of non-tree-like 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* (e.g., (Huson and Bryant, 2006)) are graphical models that capture incompatibilities in the data due to various factors, not necessarily HGT or hybrid speciation. The second category is that of *recombination networks* (e.g., (Gusfield and Bansal, 2005)), which are used to model the evolution of haplotypes and genes at the population level. *HGT 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); We henceforth refer by *phylogenetic networks* to the latter type. See (Linder *et al.*, 2004) for a detailed survey of the various phylogenetic network models and methodologies.

One of the most accurate and commonly used criteria for reconstructing phylogenetic trees is *maximum likelihood* (ML) (Felsenstein, 1981). Roughly speaking, this criterion considers a phylogenetic tree from a probabilistic perspective as a generative model, and seeks the model (i.e., tree) that maximizes the likelihood of observing a given set of sequences at the leaves of the tree.

Likelihood in the general network setting has been investigated in the past by various works. However, no HGT-specific likelihood framework has ever been suggested. von Haeseler and Churchill (Haeseler and Churchill, 1993) provided a framework for evaluating likelihoods on networks and subsequently (Strimmer and Moulton, 2000) provided an approach to assess this likelihood. These works consider a network as an arbitrary set of splits and therefore fall into the first category. They are characterized by the *combined analysis* approach, which entails combining all gene datasets first (by sequence concatenation), and then analyze the combined data set. A serious drawback of this approach is that when individual genes are governed by different evolutionary mechanisms and models (a scenario that is very common in reticulate evolution), combining multiple data sets is problematic (Nakhleh

* Authors in alphabetical order

et al., 2003). Likelihood on networks has also been considered in the setting of recombination networks (e.g., (Husmeier and McGuire, 2002)). These methods, similarly to ours, are tailored to identify breakpoints along the given sequences, however, their underlying model is different than ours as they model a different process.

In this work, we extend the ML criterion to handle specifically HGT-oriented phylogenetic networks, and propose a set of criteria and efficient heuristics for computing them. Our extension is based on the fundamental observation that, barring recombination, the evolutionary history of a gene is modeled by a tree, such that a phylogenetic network can be modeled by its constituent trees (Nakhleh *et al.*, 2005). We propose a set of ML criteria for phylogenetic networks; these criteria differ in how the tree information is used, which variant of the ML criterion is used, and finally what input is provided. Further, we investigate the computational complexity of some of these criteria and devise a set of efficient heuristics for reconstructing and evaluating phylogenetic networks based on them. In particular, we prove that scoring the likelihood of a phylogenetic network is NP-hard in general, and provide an empirically efficient exact algorithm for the problem, relying on the notion of bi-connected components. Further, we devise an efficient branch-and-bound heuristic and EM algorithm for the problem of adding a number of HGT edges to a tree to obtain an optimal phylogenetic network.

We have implemented our criteria and heuristics and studied their performance on biological as well as synthetic data. For the biological data, we analyzed two datasets. The first dataset includes the Rubisco gene in eubacteria and plastids, which was previously analyzed by Delwiche and Palmer, who postulated a set of HGT events for it (Delwiche and Palmer, 1996). The second dataset includes ribosomal protein *rpl12e* of a group of 14 Archaeal organisms, that was suspected to include HGT events (Tailliez *et al.*, 2002).

For the synthetic data, we simulated multiple data sets with various HGT events and applied our techniques to the data. In both cases, our criteria and heuristics performed very well with respect to the identification of the correct number of HGT events as well as the their placements on the organismal trees.

2 MAXIMUM LIKELIHOOD OF PHYLOGENETIC 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 and $I(T)$ its internal nodes. Further, let χ be a set of taxa (species). Then, T is a phylogenetic tree over χ if there is a bijection between χ and $L(T)$. Henceforth, we will identify the taxa set with the leaves they are mapped to, and let $[n] = \{1, \dots, n\}$ 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 node r with in-degree 0.

A phylogenetic network $N = N(T) = (V', E')$ over the taxa set χ is derived from a rooted tree $T = (V, E)$ by adding a set H of edges to T , without creating cycles, 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, $v_{e'}$; (3) finally, add a directed *reticulation edge* from v_e to $v_{e'}$. The resulting network is a rooted directed acyclic graph.

Fig. 1(a) shows a phylogenetic network obtained by adding the edge (X, Y) to the underlying organismal tree.

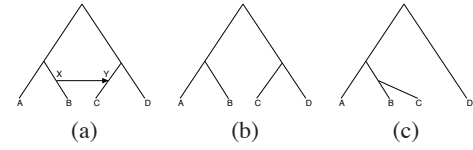


Fig. 1. (a) A phylogenetic network with a single HGT even from X to Y . (b) The underlying organismal (species) tree. (c) The tree of a horizontally transferred gene.

The tree in Fig. 1(b) models the evolution of all genetic material that is vertically inherited from the ancestral organism (the species tree), whereas the tree in Fig. 1(c) models the evolution of horizontally transferred genetic material. 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 . For example, the set $T(N)$ of the network N in Fig. 1(a) contains only the two trees which are shown in Fig. 1(b) and 1(c).

2.2 Likelihood of Phylogenetic Networks

Under the ML criterion, a phylogenetic tree is viewed as a probabilistic model from which input sequences S are assumed to be sampled. For given input sequences S the i th site, S_i , is the set of values at the i th position for every sequence in S ¹. In this work we assume that the sites are independently and identically distributed (iid). Since the parameters of the phylogenetic tree, M , are unknown, they are usually estimated from the observed sequences by maximizing the likelihood function $P(S|M)$ (Felsenstein, 1981). In general, the overall likelihood of the aligned sequences S given the model M is obtained by the product of the likelihood of every site i given M as follows:

$$L(S|M) = \prod_{i=1}^k L(S_i|M), \quad (1)$$

where k is the sequence length. Therefore, unless explicitly indicated, we will consider the likelihood of a single

¹ can be viewed as the i th column when the sequences are aligned

site henceforth. The most likely model is the one maximizing Equation 1.

When calculating the likelihood of a tree per a given site, two variants are considered: *average likelihood* (Steel and Penny, 2000) and *ancestral likelihood* (Pupko et al., 2000) (the former is the more popular of the two). The ML criterion assumes a model of evolution. We consider here the Jukes-Cantor model of sequence evolution (Jukes and Cantor, 1969). However, all the results here can be easily generalized to any other group-based model of sequence evolution. Our concept and major parts of our results can also be generalized to cases when independence across edges is not maintained. Given a set of aligned sequences $S \in \Sigma^{n \times k}$, a tree T with $I(T)$ internal nodes ($|I(T)| \leq n - 2$), and the edge transition probabilities $\mathbf{p}$, $L_{av}(S_i|T, \mathbf{p})$, the average likelihood of obtaining site S_i under T is defined as

$$L_{av}(S_i|T, \mathbf{p}) = \sum_{\mathbf{a} \in \Sigma^{|I(T)|}} \prod_{e \in E(T)} m(p_e, S_i, a), \quad (2)$$

where a ranges over all combinations of assigning labels to the $I(T)$ internal nodes of T . Each term $m(p_e, S_i, a)$ is either $p_e/(|\Sigma| - 1)$ or $(1 - p_e)$, depending on whether in the i -th site of S and a , the two endpoints of e are assigned different character states (and then $m(p_e, S_i, a) = p_e/(|\Sigma| - 1)$) or the same character state (and then $m(p_e, S_i, a) = 1 - p_e$). In the other variant of likelihood for phylogenetic trees, the ancestral likelihood, we replace the summation in Equation 2 with maximization as follows:

$$L_{anc}(S_i|T, \mathbf{p}) = \max_{\mathbf{a} \in \Sigma^{|I(T)|}} \prod_{e \in E(T)} m(p_e, S_i, a). \quad (3)$$

That is, we seek a unique labeling to internal nodes to maximize the expression. The ML solution (or solutions) for a specific tree T is the point (or points) in the edge space $\mathbf{p} = [p_e]_{e \in E(T)}$ that maximizes the expression $L(S|T, \mathbf{p})$ of Equations 2 and 3, where $M = (T, \mathbf{p})$. We will refer to both these criteria when evaluating the likelihood of a network.

A natural way of extending this setting to networks is as follows. The topology of a phylogenetic network is defined as above, however in this case since tree edges have transition probabilities, when adding a *reticulation edge* between the edges $e, e' \in E$ we should mention where along the edges e , and e' we add the two new vertices. The transition probability of the reticulation edge is always 0, meaning there are no substitutions along it (the reason being that HGT is instantaneous at the scale of evolution). However each reticulation edge $r = (e, e')$ has reticulation probability b_r associated with it. This probability denotes the probability of a DNA segment being transferred along that edge. Fig. 2 describes a simple phylogenetic network.

Let $re(T)$ denote the set of reticulation edges used to obtain tree T in the network N , and let $H(N)$ denote the set of all

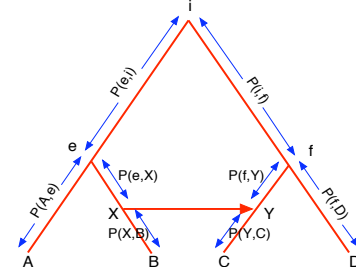


Fig. 2. Simple example of phylogenetic network under the likelihood setting.

reticulation edges in N . Let

$$P(T) = \prod_{r \in re(T)} b_r \prod_{r \in H(N) \setminus re(T)} (1 - b_r),$$

where $\mathbf{b}$ denotes the reticulation edge probabilities.

The likelihood of a network is obtained as a function of the likelihoods of the trees contained in it. Here again we consider two variants. In the first, the likelihood is the sum of the likelihood of **all** the trees of the network, where for each tree T we also need to multiply the resultant likelihood by $P(T)$. Again, we can choose between the two tree likelihood functions described above (Equations 2 and 3). Thus we get the following equation:

$$L^{all}(S_i|N, \mathbf{p}, \mathbf{b}) = \sum_{T \in \mathcal{T}(N)} P(T) \cdot L(S_i|T, \mathbf{p}) \quad (4)$$

In the other variant we want to reconstruct the sequence of reticulation events. Thus we want to find for each site one tree such that the likelihood of the leaf labels is maximized; we get the following equation:

$$L^{best}(S_i|N, \mathbf{p}, \mathbf{b}) = \max_{T \in \mathcal{T}(N)} (P(T) \cdot L(S_i|T, \mathbf{p})) \quad (5)$$

We stress that a tree likelihood can be computed only once the network is given as it is a component in the network. In order to complete the definition of the maximum likelihood of phylogenetic networks, we add the last criterion which is the type of the input provided. Therefore, we can define multiple ML criteria, depending on three issues (total of 12 variants):

1. **Likelihood criterion for the trees:** Is ancestral likelihood or average likelihood used to assess each tree likelihood in the network?
2. **Tree criterion:** At each site, is the best tree likelihood or the sum of the likelihoods over all trees taken?
3. **Input data:** Which of the network's parameters (topology and/or probabilities) are given? We consider three possibilities:
 - a. The *tiny* problem: The network topology, transition probabilities and reticulation probabilities are given.
 - b. The *small* problem: The initial tree and reticulation edges (i.e. network topology) are given but not the transition or reticulation probabilities.

- c. The *big* problem: An initial network (usually a tree) is given and a set of additional reticulation edges is sought.

3 ALGORITHMS

In the supplementary material for the paper (see the abstract) we give proofs for the NP-hardness of part of the tiny and small versions of the problems. We show that the following problems are NP-hard: Tiny best tree ancestral sequences, tiny best tree average sequences, tiny all trees ancestral sequences, tiny all trees average sequences, and small best tree ancestral sequences. In this section we describe algorithms and heuristics for dealing with these NP-hard problems. The simplest algorithm for the tiny problem is to decompose the network into all its constituent trees and analyze each tree separately by either the algorithm of (Felsenstein, 1981) for the average case or of (Pupko *et al.*, 2000) for the ancestral case. The complexity of such a naive approach is $O(2^r n)$ for each site, where r is the number of reticulation edges in the network and n is the number of leaves in the tree.

The Component-Wise Naive Algorithm for the Tiny Problem. We now describe a more efficient algorithm that takes into consideration independent components in the network. For a network N and a node $v \in V(N)$, N_v denotes the set of nodes that are reachable from v . Let u, v be two nodes in N . We say that u and v are *unrelated* if $u \notin N_v$ and $v \notin N_u$. In order to compute likelihood in a bottom-up fashion, we need the following property to hold: for every two internal nodes u, v , $(N_u \subseteq N_v)$, $(N_v \subseteq N_u)$, or $N_u \cap N_v = \emptyset$. Note that this property always holds for trees, but not necessarily for networks. Indeed, this is the underlying principle in Felsenstein's pruning algorithm to compute likelihood on a tree (Felsenstein, 1981).

DEFINITION 3.1. A biconnected component (*bi-component*) is a subgraph induced by a maximal set of vertices W , such that in the underlying graph, there are two vertex disjoint paths between any two vertices in W (we assume all the edges are undirected).

A node in a bi-component B is a *leaf* in B if it has no children in (the directed graph of) B . Otherwise it is *internal* in B . Also, a node is a *root* in B if it has no ancestor in B . It is easy to see that every bi-component has at least one leaf and exactly one root. Also, every two bi-components are internal-vertex-disjoint.

OBSERVATION 3.2. Let B be a bi-component with a root r . Let $V(B)$ denote the set of vertices of B . Then for every internal node $u \in V(B) \setminus \{r\}$ there is an unrelated internal node $v \in V(B)$ s.t. $N_u \cap N_v \neq \emptyset$ (and by definition $N_u \not\subseteq N_v$). In particular, every leaf in a bi-component has at least two parents.

The above observation implies that a more efficient algorithm than the naive algorithm can be devised to handle bi-components independently.

DEFINITION 3.3. Let N be a network. Then $B(N)$, the bi-component graph of N is a graph $(V(B), E(B))$ where V is the set of bi-components in N and two bi-components B_1, B_2 are connected by a (directed) edge in $B(N)$ if 1. There is an edge in N between $v_1 \in B_1$ and $v_2 \in B_2$. 2. There is a vertex $v \in V(B_1) \cap V(B_2)$, and v is a leaf in B_1 (and necessarily internal in B_2).

We have the following observation.

OBSERVATION 3.4. Let N be a phylogenetic network. Then $B(N)$ is a tree.

Based on the above observation, a better solution is to decompose the network into bi-components, run the naive exhaustive algorithm inside every bi-component and the tree algorithm in the bi-component. Let $r(B)$ be the number of reticulation edges in a bi-component B , B^* the largest bi-component in N and r^* the maximum of $r(B)$ over all bi-components of N . Then, the above improvement reduces the complexity of the algorithm to $O(|B(N)|2^{r^*}|B^*|)$. This improvement can be quite important as the bi-components can be sparse in a network with few reticulation edges.

Heuristics for the Small Versions. In this case, we have the network topology but not the probabilities, which we infer from the data. For the average likelihood version, we use hill climbing to compute the optimal parameters for the given topology. For the ancestral likelihood version we propose a new Expectation—Maximization (EM) algorithm. We start with random initial edge lengths. Next we perform the following two steps until convergence: (1) Find optimal internal assignments and a tree topology at each site, given the edge lengths; (2) Find the best edge lengths, given these assignments and tree topologies. Step (1) can be performed using our exact algorithm for the tiny problem. Step (2) can be performed in polynomial time as we now describe. After step (1) we have labeling for each node in the tree, and a tree for each site. For an edge e , let S_e denote the set of sites (out of k sites) where the edge e appears in the best tree. Let $h(S_e, e)$ denote the Hamming distance between the two endpoints of the edge for the sites in S_e . In step (2) we set $p_e = \frac{h(S_e, e)}{|S_e|}$, for tree edges e , and $b_e = \frac{|S_e|}{k}$, for reticulation edges e . It can be shown that repeating this procedure leads to a local critical point on the likelihood surface.

Heuristics for the Big Versions. For accelerating the running time for the big problems, we use a branch and bound (B&B) heuristic. In general, the B&B technique is based on the decreasing monotonicity of the objective function with respect to partial inputs. In other words, the value of the objective function for a certain input is not greater than any valid part of that input. The B&B principle asserts that if the value of the objective function for some partial input is smaller than some known value on the total input, then exploring all inputs that extend this partial input can be avoided. B&B is normally applied to NP-hard problem and can lead to very efficient running times. In our case, the input is a phylogenetic

network and a set of characters and the objective function is the likelihood of the network w.r.t. the character set.

Let $N = (V, E)$ be a phylogenetic network. We say that $N' = (V', E')$ is an *edge separated* subnetwork of N if $V' \subseteq V$, $E' = \{(u, v) \in E : u, v \in V'\}$, and there exists a single edge that connects a node in V and a node in V' . The following observations establish an upper bound for the likelihood of a phylogenetic network.

OBSERVATION 3.5. *If a network N' is a valid phylogenetic network and is an edge separated subnetwork of phylogenetic network N , then $P(S|N') > P(S|N)$.*

For a set S of sequences of length k , we denote by S_n^m ($n \geq 0$, $m \leq k$) the set of these sequences restricted only to the positions $n \leq i \leq m$.

OBSERVATION 3.6. *For any phylogenetic network, N and set S of sequences, we have $P(S_n^m|N) \geq P(S|N)$.*

Based on these observations, we propose a 2-step heuristic: (1) Optimize the likelihood of sub-networks; if a candidate network contains an edge separated sub-network with low likelihood score, it is not the network with the ML score; (2) Optimize the likelihood of a network on part of the sites; if a candidate network has low likelihood score for these sites, it is not the network with the ML score.

4 EMPIRICAL PERFORMANCE

We implemented our ML criteria and algorithms and tested them first on simulated data. Next, using insights from this study we applied our software on two real biological data sets. The problem we sought to solve is the *big* problem in which only the organismal tree and the input sequences are given and the task is to reconstruct the network topology and sets of edge probabilities. Specifically we investigated the performance of ML with respect to its ability to infer the correct number and location of HGT events.

Simulation. We used the *r8s* tool 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. We then augmented this tree with five reticulation edges (simulating HGT events). Next we used the Seq-gen tool (Rambaut and Grassly, 1997) to evolve DNA sequences on the resulting network: The first 1500 sites down the organismal tree, and subsequent 500 sites down the tree obtained by applying the five transfer events to the organismal tree. Both sequence datasets were evolved under the K2P+ γ model of evolution, with shape parameter 1 (Kimura, 1980). The best results we obtained for this type of experiment were under the ancestral likelihood model. In all our experiments (26) the best reconstructed edges (the ones that contributed the most to the network likelihood) were indeed the HGT edges (in the model network). The sixth edge exhibited a lesser contribution and the seventh lesser than that. Moreover, all the

sites where the HGT edges were inferred (but not all of them) were between the 1500th and the 2000th sites.

Results on Biological Data. We analyzed two biological data sets. In the first we considered the rubisco gene *rbcL* of 15 plastids, cyanobacteria, and proteobacteria organisms. This is a subset of the dataset considered by Delwiche and Palmer (Delwiche and Palmer, 1996) (due to ML computational intensity, we could not analyze the whole 48-taxon set) for which multiple HGT events were conjectured by the authors. The dataset consists of two sequences from each of the α , β , and γ -proteobacteria groups, 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 (organismal) tree which was reported in (Delwiche and Palmer, 1996; Boc and Makarenkov, 2003). The species tree is based on 16S rRNA and other evidence. The 532 sites long alignment is available from <http://www.life.umd.edu/labs/delwiche/alignments/rbcLgb7-95.distrib.txt>.

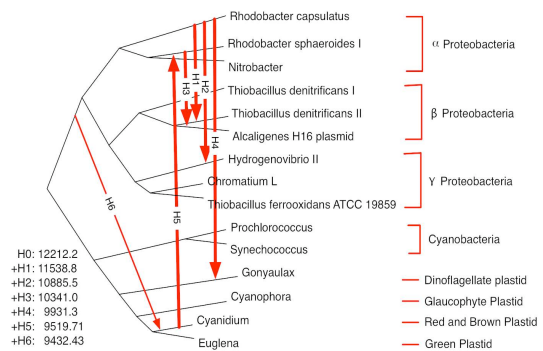


Fig. 3. The species tree of the 15 organisms, as reported by Delwiche and Palmer, and the 6 HGT edges inferred by our heuristic for the big ML problem. Likelihood criterion = *ancestral*, and tree criterion = *all*.

Delwiche and Palmer hypothesized the occurrence of several HGT events of the rubisco genes as opposed to ancient gene duplication and loss. Our findings, based on the ML criterion are as follows. (1) The two most significant HGT edges are H1 and H2 in Fig. 3, and they group the form II rubisco (Thiobacillus denitrificans II and Hydrogenovibrio II) of the β and γ proteobacteria, respectively, together with the form II rubisco (Rhodobacter capsulatus) of the α proteobacteria. This result agrees with the grouping of these three organisms, based on the form II rubisco, into one clade, as shown in Fig. 2 of (Delwiche and Palmer, 1996). (2) The third most significant HGT edge is H3 in Fig. 3, and it indicates an HGT of the red type form I rubisco from the α proteobacteria (Rhodobacter sphaeroides I) to the β proteobacteria (Alcaligenes H16 plasmid). This result agrees with the grouping of all red type form I rubisco genes in one clade in Fig. 2 of (Delwiche

and Palmer, 1996). (3) The fourth most significant HGT edge is H4 in Fig. 3, which completes the grouping of the form II rubisco genes, along with H1 and H2, by indicating an HGT of the form II rubisco from *Rhodobacter capsulatus* to *Gonyaulax* (an α proteobacteria and plastid, respectively). This result is also supported by the single clade of all form II rubisco in Fig. 2 of (Delwiche and Palmer, 1996). (4) The fifth most significant HGT edge is H5 in Fig. 3, which indicates an HGT of the form I rubisco from the red and brown plastids (*Cyanidium*) to the α proteobacteria (*Rhodobacter sphaeroides* I). This HGT event is in agreement with the grouping of the red type form I rubisco in Fig. 2 of (Delwiche and Palmer, 1996). (5) The last HGT edge is H6 in Fig. 3 which groups the red and brown plastids with the α , β , and γ proteobacteria. Such grouping is supported by the red type form I rubisco group in Fig. 2 of (Delwiche and Palmer, 1996). To summarize, the HGT edges computed by our heuristic agree with the grouping of the organisms based on the forms I and II rubisco, as hypothesized by Delwiche and Palmer.

We also compared our results to the results reported by Boc and Makarenkov (Boc and Makarenkov, 2003), who used a distance method for discovering HGT and analyzed a dataset similar to ours. Three of our edges (H_1 , H_3 , H_4) appeared in (Boc and Makarenkov, 2003). Two other edges (H_2 , H_5) appeared also in (Boc and Makarenkov, 2003), but in the opposite direction. One edge, H_6 , appeared in our results but didn't appear in the results of (Boc and Makarenkov, 2003). In general, our results are similar to the result reported by Boc and Makarenkov, but simpler. Our results included 6 HGTs, while the solution of Boc and Makarenkov included 8 HGTs.

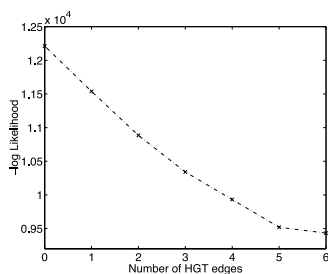


Fig. 4. The improvement in the likelihood score as a function of the number of HGT edges added to the species tree (shown in Fig. 3). The improvement achieved by adding the sixth HGT edge is smaller compared to that achieved by adding the first five events, which indicates that five HGT events suffice to model the evolution of the *rbcl* gene for these 15 organisms. Likelihood criterion = *ancestral*, and tree criterion = *all*.

The second biological dataset includes the ribosomal protein *rpl12e* over a group of 14 archaeal organisms, that was mentioned in the work of Tailliez *et al.* (Tailliez *et al.*, 2002). We used the organismal tree described in (Tailliez *et al.*, 2002) that was based both on the concatenation of 57 ribosomal proteins (7,175 positions), and on the concatenation of SSU and

LSU rRNA (3,933 positions) (the two methods gave identical trees). The sequences were downloaded from NCBI and were aligned by ClustalW. Tailliez *et al.* (Tailliez *et al.*, 2002) claimed that the phenogenetic tree based on these proteins is different than the organismal tree of these archaea. We used our method for explaining this difference. The results are described in figures 5 and 6 and suggest that three HGT events can explain the difference between the organismal tree and the phylogenetic tree for the *rpl12e* proteins.

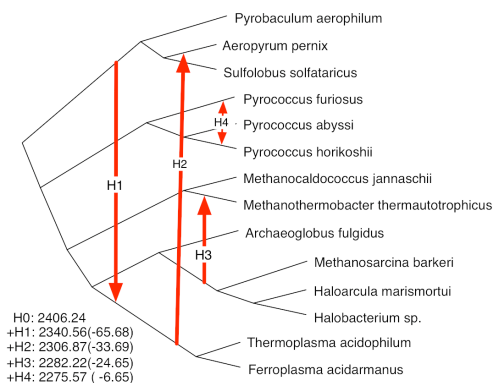


Fig. 5. The species tree of the 14 organisms, as reported by Tailliez *et al.*, and the 4 HGT edges inferred by our heuristic for the big ML problem. Likelihood criterion = *ancestral*, and tree criterion = *all*.

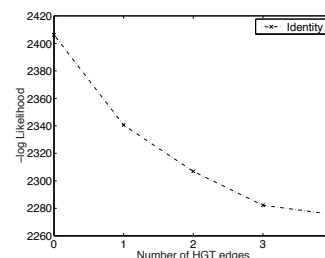


Fig. 6. The improvement in the likelihood score as a function of the number of HGT edges added to the species tree (shown in Fig. 5). The improvement achieved by adding the fourth HGT edge is smaller compared to that achieved by adding the first three events, which indicates that three HGT events suffice to model the evolution of the ribosomal protein *rpl12e* gene for these 14 archaea. Likelihood criterion = *ancestral*, and tree criterion = *all*.

From the definition of the ML criterion for networks it follows that adding an edge never decreases the likelihood score. Therefore, a significant question is when to stop adding edges (stopping rule). To answer this question, we plotted the improvement in the likelihood score as a function of the number of HGT edges added; the results are shown in Fig. 4 for the first data set and Fig. 6 for the second dataset. The first figure shows that while adding the first five edges achieves drastic improvements in the likelihood score, adding the sixth

edge results in a much lower improvement, which is indicated by the slow decrease in the likelihood score. Similarly, the second figure shows that the first three edges achieve the major improvements in the likelihood score.

5 CONCLUSIONS AND FUTURE RESEARCH

Phylogenetic networks model evolutionary histories of sets of organisms in the presence of non-treelike evolutionary events such as HGT and hybrid speciation. In this paper, we introduced a new maximum likelihood framework for reconstructing and evaluating phylogenetic networks. This framework gave rise to an array of computational problems. We addressed the most basic and fundamental of these problems such as the complexity of reconstruction, hardness of the “tiny” variants, devising efficient heuristics and algorithms, and showing the viability of this criterion. We implemented our methods and tested them on a large set of simulated data. We also analyzed two biological data sets. On the dataset of eubacteria and plastids (Delwiche and Palmer, 1996) we confirmed previous conjectures made by Delwiche and Palmer, and on the archaeal dataset of (Tailliez *et al.*, 2002) we explained the discrepancy detected by Tailliez *et al.* by three HGT edges. The empirical analysis done here was aimed to demonstrate the relevance of the ML criterion. However, as even the simplest models we assumed here are NP-hard, it is easy to be convinced that even the relatively simple methods we employed here are fairly involved.

Future research directions include (1) developing more computationally efficient algorithmic techniques to enable analysis of large data sets; (2) modeling of dependence among sites (we intend to investigate HMMs for this purpose; and (3) exploring various distributions of reticulation edge probabilities. It is clear that there is a tradeoff between the first and the other two points. While using HMMs and more complicated distributions of reticulation probabilities will make the model more accurate, it will also dramatically increase the running time of the method. Thus we believe that simpler probabilistic models, as we described here, are practically important.

ACKNOWLEDGMENTS

The authors would like to thank Derek Ruths and Satish Rao for helpful discussions and data, and the anonymous reviewers for helpful comments. Experiments were run on the Rice Terascale Cluster, funded by NSF under grant EIA-0216467, Intel, and HP. L.N. was supported in part by DOE grant DE-FG02-06ER25734 and NSF grant CCF-0622037, and S.S. was supported in part by NSF grant CCR-0105533.

REFERENCES

- A. Boc and V. Makarenkov. New efficient algorithm for detection of horizontal gene transfer events. In *WABI03*, pages 190–201, 2003.
- Husmeier D. and McGuire G. Detecting recombination with mcmc. *Bioinformatics*, **18**:345–353, 2002.
- C. F. Delwiche and J. D. Palmer. Rampant horizontal transfer and duplication of rubisco genes in eubacteria and plastids. *Mol. Biol. Evol.*, **13**(6), 1996.
- W.F. Doolittle, Y. Boucher, C.L. Nesbo, C.J. Douady, J.O. Andersson, and A.J. Roger. How big is the iceberg of which organellar genes in nuclear genomes are but the tip? *Phil. Trans. R. Soc. Lond. B. Biol. Sci.*, **358**:39–57, 2003.
- I.T. Paulsen *et al.* Role of mobile DNA in the evolution of *Vacomycin-resistant Enterococcus faecalis*. *Science*, **299**(5615):2071–2074, 2003.
- J. Felsenstein. Evolutionary trees from DNA sequences: A maximum likelihood approach. *J. Mol. Evol.*, **17**:368–376, 1981.
- D. Gusfield and V. Bansal. A fundamental decomposition theory for phylogenetic networks and incompatible characters. In *RECOMB05*, pages 217–232, 2005.
- D. H. Huson and D. Bryant. Application of phylogenetic networks in evolutionary studies. *Mol. Biol. Evol.*, **23**(2):254–267, 2006.
- T. H. Jukes and C. R. Cantor. Evolution of protein molecules. in mammalian-protein metabolism. *H. N. Munro(Ed)*, 121-132 *New York: Academic Press*, 1969.
- M. Kimura. A simple method for estimating evolutionary rates of base substitutions through comparative studies of nucleotide sequences. *J. Mol. Evol.*, **16**:111–120, 1980.
- C.R. Linder, B.M.E. Moret, L. Nakhleh, and T. Warnow. Network (reticulate) evolution: biology, models, and algorithms. *PSB04. A tutorial*.
- B.M.E. Moret, L. Nakhleh, T. Warnow, C.R. Linder, A. Tholse, A. Padolina, J. Sun, and R. Timme. Phylogenetic networks: modeling, reconstructibility, and accuracy. *TCBB*, **1**(1):13–23, 2004.
- L. Nakhleh, D. Ruths, and H. Innan. Gene trees, species trees, and species networks. In R. Guerra and D. Allison, editors, *Meta-analysis and Combining Information in Genetics*. Chapman & Hall, CRC Press, 2005. In press.
- L. Nakhleh, J. Sun, T. Warnow, C.R. Linder, B.M.E. Moret, and A. Tholse. Towards the development of computational tools for evaluating phylogenetic network reconstruction methods. In *PSB03*, 2003.
- T. Pupko, I. Pe’er, R. Shamir, and D. Graur. A fast algorithm for joint reconstruction of ancestral amino-acid sequences. *Mol. Biol. Evol.*, **17**(6):890–896, 2000.
- A. Rambaut and N. C. Grassly. Seq-gen: An application for the Monte Carlo simulation of DNA sequence evolution along phylogenetic trees. *Comp. Appl. Biosci.*, **13**:235–238, 1997.
- M. Steel and D. Penny. Parsimony, likelihood and the role of models in molecular phylogenetics. *Mol. Biol. Evol.*, **17**:839–850, 2000.
- K. Strimmer and V. Moulton. Likelihood analysis of phylogenetic networks using directed graphical models. *Mol. Biol. Evol.*, **17**(6):875–881, 2000.
- O.M. Tailliez, C. Brochier, P. Forterre, and H. Philippe. Archaeal phylogeny based on ribosomal proteins. *Mol. Bio. Evol.*, **19**(5):631–639, 2002.
- A. von Haeseler and G. A. Churchill. Network models for sequence evolution. *J. Mol. Evol.*, **37**:77–85, 1993.