

Novel Phylogenetic Network Inference by Combining Maximum Likelihood and Hidden Markov Models (Extended Abstract)

Sagi Snir¹ and Tamir Tuller²

¹ Dept. of Mathematics and Computer Science
Netanya Academic College, Netanya
ssagi@Math.Berkeley.EDU

² School of Computer Science
Tel Aviv University
tamirtul@post.tau.ac.il

Abstract. Horizontal Gene Transfer (HGT) is the event of transferring genetic material from one lineage in the evolutionary tree to a different lineage. HGT plays a major role in bacterial genome diversification and is a significant mechanism by which bacteria develop resistance to antibiotics. Although the prevailing assumption is of complete HGT, cases of partial HGT (which are also named *chimeric* HGT) where only part of a gene is horizontally transferred, have also been reported, albeit less frequently.

In this work we suggest a new probabilistic model for analyzing and modeling phylogenetic networks, the NET-HMM. This new model captures the biologically realistic assumption that neighboring sites of DNA or amino acid sequences are *not* independent, which increases the accuracy of the inference. The model describes the phylogenetic network as a Hidden Markov Model (HMM), where each hidden state is related to one of the network's trees. One of the advantages of the NET-HMM is its ability to infer partial HGT as well as complete HGT. We describe the properties of the NET-HMM, devise efficient algorithms for solving a set of problems related to it, and implement them in software. We also provide a novel complementary significance test for evaluating the fitness of a model (NET-HMM) to a given data set.

Using NET-HMM we are able to answer interesting biological questions, such as inferring the length of partial HGT's and the affected nucleotides in the genomic sequences, as well as inferring the exact location of HGT events along the tree branches. These advantages are demonstrated through the analysis of synthetical inputs and two different biological inputs.

1 Introduction

Eukaryotes evolve largely through vertical lineal descent in a tree-like manner. However, in the presence of HGT the right model of evolution is not a tree but is rather 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 [16]. In the case of a complete HGT the assumption is that a single tree (one of the trees induced by the network) describes the evolution of a gene, while in the case of a chimeric HGT more than one tree is needed (*i.e.* different parts of a gene evolve according to different trees).

HGT (partial or complete) is very common among *Bacteria* and *Archaea* [6,8,5], but evidence of HGT (partial or complete) between Eukaryotes are also accumulating [2,26]. A large body of work has been introduced in recent years to address phylogenetic network reconstruction and evaluation. Methods for dealing with this problem include a variety of approaches: *Splits Networks* (see *e.g.* [14]) which are graphical models that capture incompatibilities in the data due to various factors, not necessarily HGT or hybrid speciation; Maximum Parsimony (MP) [12,16,17] that are based on Occam's Razor approach; distance methods, that try to fit a distance matrix to a network [4] or use the minimum evolution criterion [3], and graph theoretical approaches that try to fit a gene tree to a species tree [1,10,23,11].

One of the most accurate and commonly used criteria for reconstructing phylogenetic trees is *Maximum Likelihood* (ML) [9]. 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 the given input set of sequences at the leaves of the tree. Likelihood in the general network setting has been investigated in the past in various studies. Von Haeseler and Churchill [29] provided a framework for evaluating likelihood on networks and subsequently [28] provided an approach to assess this likelihood. These works consider a network as an arbitrary set of splits and do not correspond to a specific biological process. Likelihood on networks has also been considered in the setting of recombination networks (see *e.g.* [13]). These methods are tailored to identify breakpoints along the given sequences. However, their underlying model, the biological questions they investigate, and the algorithmic approaches they pursue are different from ours as they model a different biological process.

Recently, Jin *et al.* performed an initial step toward developing an HGT-oriented likelihood based model for evolutionary networks [16]. This work demonstrated the potential of using ML for inferring evolutionary networks. The main advantage of that work is its simplistic underlying model, that enables efficient implementation. Another related work is the study of Siepel and Haussler [27] who suggested a model that combines a phylogenetic tree along with a HMM and used it for aligning full genomes. Our work was inspired by these two works. However, while the main goal in [16] was to infer complete HGT events, here we focus on analyzing chimeric HGT events by adopting a more biologically relevant model for this task, the *NET-HMM* model.

The NET-HMM models a phylogenetic network by a Hidden Markov Model (HMM), where each of the network's trees corresponds to a state of the HMM (see figure 1); *i.e.* the emission probability in each state is according to its corresponding tree.

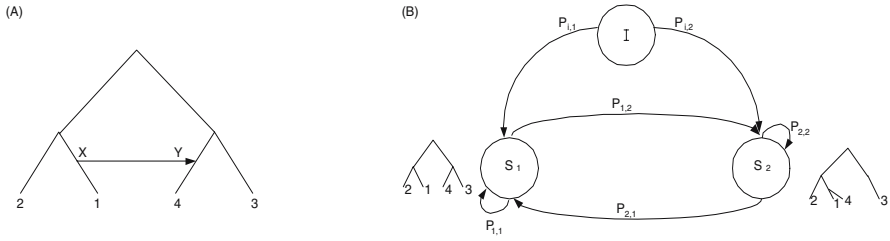


Fig. 1. Simple example of the NET-HMM model. (A) A phylogenetic network, with a single HGT from X to Y . (B) The HMM for the network: I denotes the initial state, all the other states are related to the networks' trees. The state S_1 is related to the underlining organismal tree. The state S_2 is related to the tree of realizing the horizontal transfer edge. We assume $P_{i,i} > P_{i,j}$, *i. e.* in each position the probability to stay in the same state/tree is higher than the probability of transition to other tree.

The model is supported by biological facts that (i) adjacent sites are not independent with respect to HGT, (ii) events of HGT are relatively rare. In the view of the NET-HMM model, events such as chimeric HGT [22] are described as a transition between states (trees) of the HMM. NET-HMM also reconstructs the exact HGT location on the tree edges, a task that methods such as MP or distance based methods [17,4] can not accomplish.

By applying the NET-HMM on synthetical data, we show very significant improvements over the i.i.d. model [16] in the ability to locate chimeric events and locations along tree edges. Results on biological data point out significant biological phenomena such as amelioration [15,25], chimeric HGT events, and occurrence of very recent HGT events. Applying the NET-HMM to data that were analyzed in the past suggest an excess of HGT inference. These findings suggest a further scrutiny.

In the statistical-algorithmic realm, we propose a novel algorithm, *EM NET-HMM*, that interweaves into the conventional EM algorithms a step of hill climbing to maximize emission probabilities, rendering a non trivial algorithmic approach. We also devised a novel permutation test to measure the fitness of a NET-HMM to a given dataset.

2 Methods

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 $F(T)$ 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 $F(T)$. 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. Let Σ denote the set of *states* (*e.g.* for DNA, $|\Sigma| = 4$). Then with each edge $e \in E$ we associate a *substitution probability* p_e indicating the probability of observing

different states at the endpoints of e . For a wide variety of evolutionary models there is an invertible transformation from the p_e to *edge length* q_e s.t. q_e 's are *additive* - applying the inverse transformation on the sum $q_\pi = \sum_{e \in \pi} q_e$ of edge lengths along a path π yields the probability p_π of observing different states at the endpoints of π . Therefore, under the mapping $\mathbf{q} : E \rightarrow \mathbb{R}$ of lengths, $T = (V, E, \mathbf{q})$ is a weighted tree (we omit \mathbf{q} when it is clear from the context). The edge length and additivity are crucial for our formulation as is explained in the sequel.

In this work we consider the Jukes-Cantor (JC) model of sequence evolution [21]. However, all the results here can be generalized to other models of sequence evolution. Under JC, the length-substitution relationship is as follows: $p_e = \frac{3}{4}(1 - e^{-4/3q_e})$ and $q_e = -\frac{3}{4} \ln(1 - \frac{4}{3}p_e)$.

For a given set of input sequences S , the i -th site, S_i , is the set of states at the i -th position for every sequence in S ¹. Under the ML criterion, a phylogenetic tree is viewed as a probabilistic model from which input sites are assumed to be sampled. The probability of obtaining a site S_i given a tree T , $L(S_i|T)$, is defined as [9]:

$$L(S_i|T) = \sum_{\mathbf{a} \in \Sigma^{I(T)}} \prod_{e \in E(T)} m(p_e, S_i, a), \tag{1}$$

where a ranges over all combinations of assigning states 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, under a , the two endpoints of e are assigned different or the same states respectively.

A phylogenetic network $N = N(T) = (V', E')$ over the taxa set χ is derived from a rooted weighted tree $T = (V, E, \mathbf{q})$ by adding a set R of reticulation edges to T , where each edge $r \in R$ is added as follows: (1) split an edge $e \in E$ by adding a new node, v_e , s.t. the lengths of the newly created edges sum to the length of e ; (2) split an edge $e' \in E$ by adding another new node, $v_{e'}$ (again by preserving lengths); (3) finally, add a directed *reticulation edge* r from v_e to $v_{e'}$. We add that the substitution probability (and hence the length) of r is zero as these events are instantaneous in time. The mathematical implication of the above is that it extends the partial order induced by T on V (see [19,18]). This results in having no cycles in which tree edges are traversed along their directionality and reticulation edges in either direction (see Figure 3 A.). A tree T' is induced by N by removing all but one incoming edges to the newly added nodes, and contracting degree-2 nodes (while summing the edge lengths). We denote by $T(N)$ the set of trees induced by N (see Figure 3 A.).

2.2 From a Network to a HMM

The i.i.d Model. Under the i.i.d model, each reticulation edge r has a probability denoting the corresponding event's probability. Under this formulation, every edge (including tree edges) is assigned an *occurrence* probability s.t. the

¹ Can be viewed as the i -th column when the sequences are aligned.

sum of occurrence probabilities of edges entering a node is 1. For a tree T , let E_T denote the set of (reticulation + tree) edges realized by (or giving rise to) T , and let $P(E_T)$ be the probability of observing E_T (It can be seen that this is the product of their individual probabilities).

Therefore, the likelihood of obtaining a site, S_i , given a phylogenetic network, N , is [16]: $L(S_i|N) = \sum_{t \in T(N)} P(E_t) \cdot L(S_i|t)$. The likelihood of obtaining the input sequences is $L(S|N) = \prod_i L(S_i|N)$. As different sites are modeled independently, a weak signal at a certain site will cause the inference of an erroneous tree at that site (a phenomenon we give more details on in the simulation section). To overcome this shortcoming, we treat the sites as a Markovian process, as we describe in the next subsection.

The NET-HMM. The NET-HMM is a tuple $M = \{N, H\}$ where $N = (V', E', \mathbf{q})$ is a phylogenetic network, and H is a Hidden Markov Model (HMM). We do not know the evolutionary history (a tree in $T(N)$) of every site in S , thus we assign a hidden state for each site in S , and an initial state, I . The hidden states correspond to the states of H and let Σ_H denote this set. Let $\Upsilon(h)$ denote the tree related to the hidden state h (the initial state is not related to a tree, so $h \neq I$). The meaning of relating the state h of the i -th site to a state of the HMM is that this site evolves on the tree $\Upsilon(h) \in T(N)$ (*i. e.* the i -th column was emitted by the tree T).

Let $p(h_{i-1} \rightarrow h_i)$ denote the transition probability between state h_{i-1} and state h_i in the HMM. The likelihood of a NET-HMM model M when observing a set S of n -long sequences, is defined as the probability of observing S evolving on M which is the sum of probabilities of all length- n paths of states from Σ_H . Thus $L(S|M)$ equals

$$\sum_{h_1^n \in (\Sigma_H)^n} p(I \rightarrow h_1) \prod_{i=2}^n p(h_{i-1} \rightarrow h_i) \cdot L(S_i|\Upsilon(h_i), \mathbf{q}) \tag{2}$$

where h_1^n is a sequence of n states (and $\forall_i h_i \in \Sigma_H$).

A different variant of the likelihood function scores a network by the probability of the most likely path, \hat{h}_1^n , in M . The latter is achieved by replacing the sum by a maximum relation.

Our goal is to find the model (network topology, edge length, and transition probabilities of the HMM) that maximizes the likelihood of the input sequences (equation 2). By using an HMM we gain two important advantages: 1) We gain dependencies among close sites, as in reality. 2) We indirectly infer the probabilities of reticulation events, while avoiding the use of arbitrary parameters for reticulation probability (as was assumed in [16]).

We emphasize three important constraints on the NET-HMM model; these constraints are biologically motivated but also decrease the parameter space (and thus reduce the running time), while improving the quality of the results:

1. The network induces both the topology and edge lengths of its trees (see Figure 2).
2. Temporal constraints on the reticulation edges decrease the number of valid networks (see Figure 3A.).

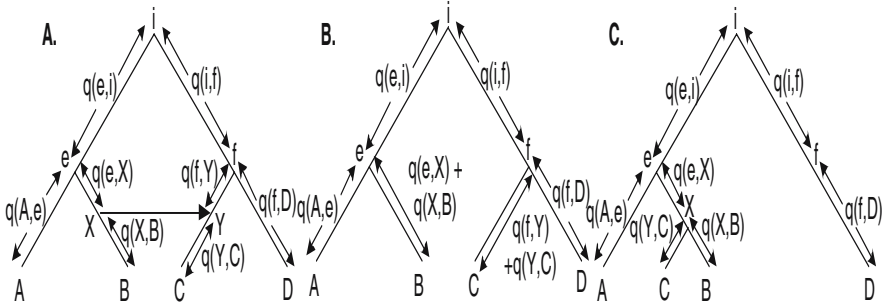


Fig. 2. A simple example of a phylogenetic network under the likelihood setting, and the set of induced trees. *A.* A phylogenetic network with one reticulation edge; each edges length denotes the expected number of subsections along the edge. *B.* One of the network’s trees that does not include the reticulation edge. *C.* One of the network’s trees that includes the reticulation edge.

3. By imposing constraints on the transition probabilities of *H*, we can drastically reduce the exponential number of paths. We add that there is a positive transition probability from the initial state to each of the other states (see Figure 3 *B.*).

Given a set of sequences associated with the tree leaves, we distinguish between three major versions of the problem, tiny, small, and large that are defined as follows:

1. **The tiny version:** Input: the network topology along with its edge lengths and transition probabilities between states in the induced HMM. Output: The most likely path in the HMM.
2. **The small version:** Input: The network topology. Output: The ML network edge probabilities and ML transition probabilities of the induced HMM.
3. **The large version:** Input: The initial organismal tree. Output: The ML NET-HMM (complete network+HMM).

The specific version of the problem should be employed depending on the given input. When only an organismal tree and a set of sequences, suspected of having undergone HGT, are given, the large problem is chosen. When some prior information about the topology of the network (i.e. a set of reticulation edges corresponding to the organismal tree) is given and we want to know the exact location of the reticulation edges (the “split” points) and the positions of the events along the sequences, we should solve the small problem. The tiny problem finds the positions of the events along the input sequences (a partial task of the small problem) by inferring the evolutionary history of each site (position) of the input sequences.

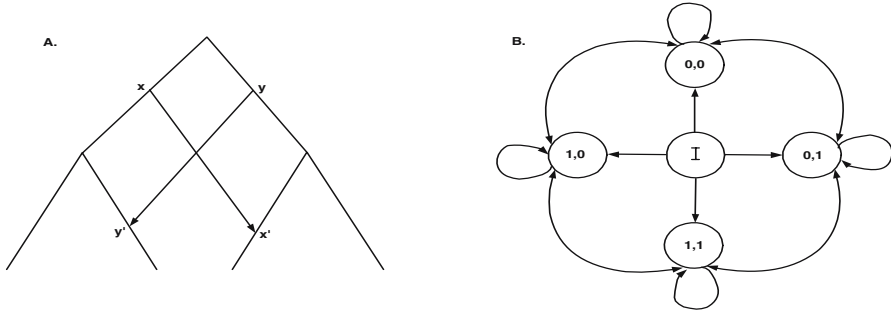


Fig. 3. Constraints on the structure of the NET-HMM. A. Example of a network that does not satisfy the temporal constraints. By the tree topology, x occurs before y' and y occurs before x' . By the reticulation edges, x and x' occur at the same time and y and y' occurs in the same time. Thus, y' occur before x' and x' occurs before y' - a contradiction. B. A simple example of the implementation of the assumption that HGTs are rare events. Suppose the phylogenetic network includes 2 edges. We name each of the network's four states/trees by its code (I denotes the initial state). The figure shows only the transitions with non-zero probabilities, *i. e.* transitions between trees (states) with hamming distance ≤ 1 .

2.3 Algorithms

The Tiny Problem - Finding the Most Likely Path. Let $L(S_1^i|h', M)$ denote the likelihood of observing the first i positions of the sequences under M restricted that the i 'th state is $h' \in H$. First we observe that $L(S|M) = \sum_{h' \in H} L(S_1^n|h', M)$. Moreover, $L(S_1^i|h', M)$ equals

$$\sum_{h'' \in \Sigma_H} L(S_1^{i-1}|h'', M) \cdot p(h'' \rightarrow h') \cdot L(S_i|\mathcal{T}(h'), \mathbf{q}). \quad (3)$$

Equation (3) is solved by the dynamic programming forward and backward algorithms [7]. The maximization variant is solved similarly by the Viterbi dynamic programming algorithm.

A related question is to infer the most likely state (tree), \hat{h}_i , at a certain site, i . For this problem we used the forward and backward algorithms and calculate: $p(S, h_i = k) = p(S_1^i, h_i = k) \cdot p(S_{i+1}^n|h_i = k)$. By the forward and backward algorithms we can calculate $p(S)$. Thus we get the results $p(h_i = k|S) = \frac{p(S, h_i=k)}{p(S)}$.

The Small Problem - Learning a Given Model. Given the network's topology (and hence the induced HMM) we use our extended EM algorithm, EM NET-HMM (depicted in Figure 4), for estimating the edge lengths and edge probabilities of the network as well as the state transition probabilities in the HMM.

EM NET-HMM:

1. Start with initial random edge lengths and transition probabilities.
2. Perform until convergence:
 - a. Given the edge lengths (that induce an emitting probability for all states); optimize the transitions probabilities by BW algorithm.
 - b. Given the transition probabilities of the HMM, optimize the edge probabilities (*i. e.* find edges that maximize the cost function that appear in equation 2) by hill climbing.

Fig. 4. The main algorithm, EM NET-HMM

Observation 1. *The algorithm EM NET-HMM terminates and converges to a local ML point.*

The Large Problem - Expanding a Given Model. Recall that the task in the large problem is to add reticulation edges s.t. the likelihood of the model is maximized. Literally, solving this problem boils down to iteratively adding edges to a given tree/network. By the fact that additional edges never decrease the likelihood of the model, some stopping criterion is required. One possibility is to observe the improvement in the likelihood score and stop when the improvement is insignificant (this approach was used in [16,17]). We however used a more rigorous approach as described in the next sub section.

2.4 Network Significance

Recall that the improvement of the NET-HMM model over the iid (see Section 3 for experimental results) was achieved by the coupling of neighboring sites. This is reflected by the transition probabilities of the HMM. Therefore in order to evaluate the significance of a given model (network+HMM) M' WRT the data order within the given input S' , we devised the following test that resembles the conventional permutation test in statistics: For a given random sampling S_0 of the sites of the input, let $L(S_0|M')$ be the likelihood of M' WRT S_0 (*i.e.* the solution of the tiny problem). For a given big enough set of such samples, we obtain a probability distribution of the likelihood of M' . Given that distribution we can compute the *empirical p-value* of M' WRT S' . In our setting, we run this test after each time a model is built, that is after the application of the small problem. We note that another significance test could be obtained by randomizing over the space of networks, however, due to the small size of this space and the computational complexity of calculating this distribution, we used only the first test.

We used this approach to determine stopping criterion. We stopped the algorithm of iteratively adding HGT edges when either the likelihood did not increase or the p-value increased.

3 Experimental Results

3.1 Synthetical Inputs

We first implemented our algorithms on synthetical data. In order to test the NET-HMM's accuracy we tried to solve the small problem on simulated data where we know the "true" model. We generated 20 synthetical phylogenetic networks, sampled them, and used these samples as inputs to our methods.

In the first test our aim was to check how well the NET-HMM reconstructs the set of events (correct positions along the correct sequences) occurred in the network. This is equivalent to finding the correct path between the HMM states (trees). We sampled a segment (a consecutive set of sites) from each of the trees in the network and concatenated them together into a basic concatenation block. We defined an error rate in the reconstruction of the most likely path as the fraction of sites where the NET-HMM inferred a wrong tree. The experiments were conducted for various HGT segment lengths and for various number of replications of the basic concatenation block. The results are shown in Figure 5.

In Figure 5.A. it can be appreciated that the NET-HMM was able to reconstruct each of the segments fairly accurately and significantly better than the i.i.d model. It demonstrates that segments of length 50 are enough for generating a surprisingly good reconstruction of the true path (error rate between 0.15 and 0.19), and segments of length 500 will give a very good reconstruction (error rate less than 0.01). The results also show that the method is independent of the number of replications of the basic concatenation block. On the other hand, as expected the error rate of the i.i.d model was around 0.8 irrespective of the number of replications and segment lengths. Figures 5.B and 5.C show the output of two typical runs of EM NET-HMM and the i.i.d. model compared with the true path, for segment lengths 50 and 500 respectively. The advantage of EM NET-HMM over the i.i.d. model is very clear and outstanding.

In our second test, we checked the ability to infer the correct HGTs by the NET-HMM (i.e. the big problem). When checking synthetic datasets of moderate sequence length (above 500 sites) and tree edges of significant length (so the correct HGT edge location is crucial) usually the improvement in the likelihood of the true HGT edges was much larger than arbitrary ones, and the NET-HMM successfully identified the correct HGTs.

3.2 Biological Inputs

As indicated above, the main strength of our method is its accuracy, making it appropriate for further analysis of results obtained or hypothesized by other methods. We believe this is the most practical way of using our model since it substantially decreases the number of learned parameters and hence the statistical and the computational complexity of the problem.

Hence, we solved a restricted version of the small problem where the task is to infer the exact position of the reticulation edges along the tree branches and the most likely tree at each site. The organismal tree (topology and edge lengths) was inferred using ML on various sets of genes and was taken as constant.

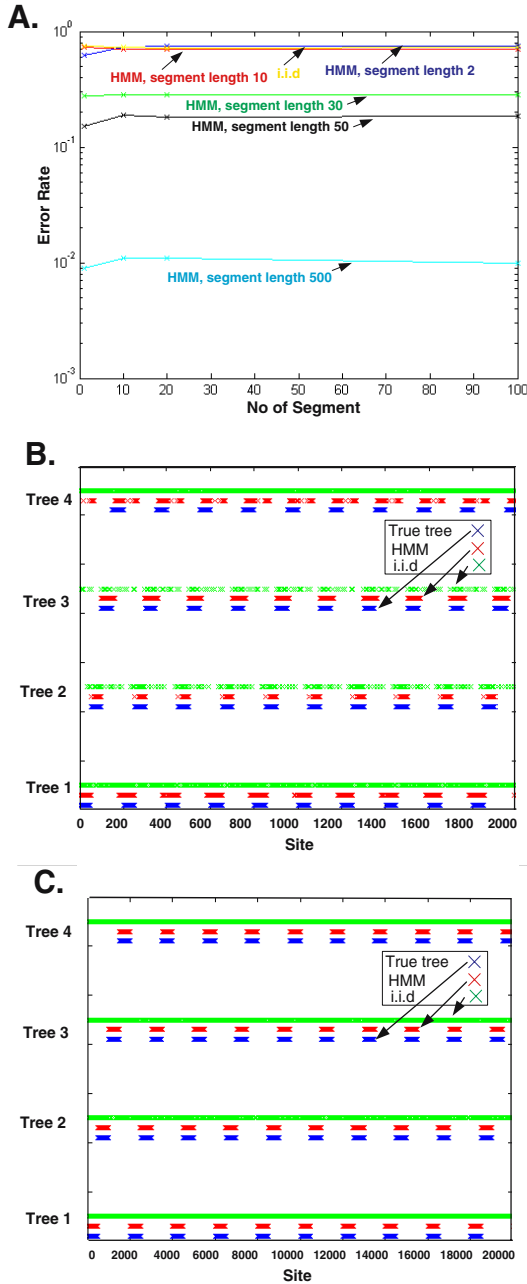


Fig. 5. The performances of the NET-HMM on synthetic inputs (a network with 8 leaves and two reticulation edges). *A.* Reconstructing the tree in each site: Error rate for different segment lengths. *B.* Segment length 50: The inferred tree in each site by NET-HMM vs the i.i.d. model. *C.* Segment length 500: The inferred tree in each site by NET-HMM vs the i.i.d. model.

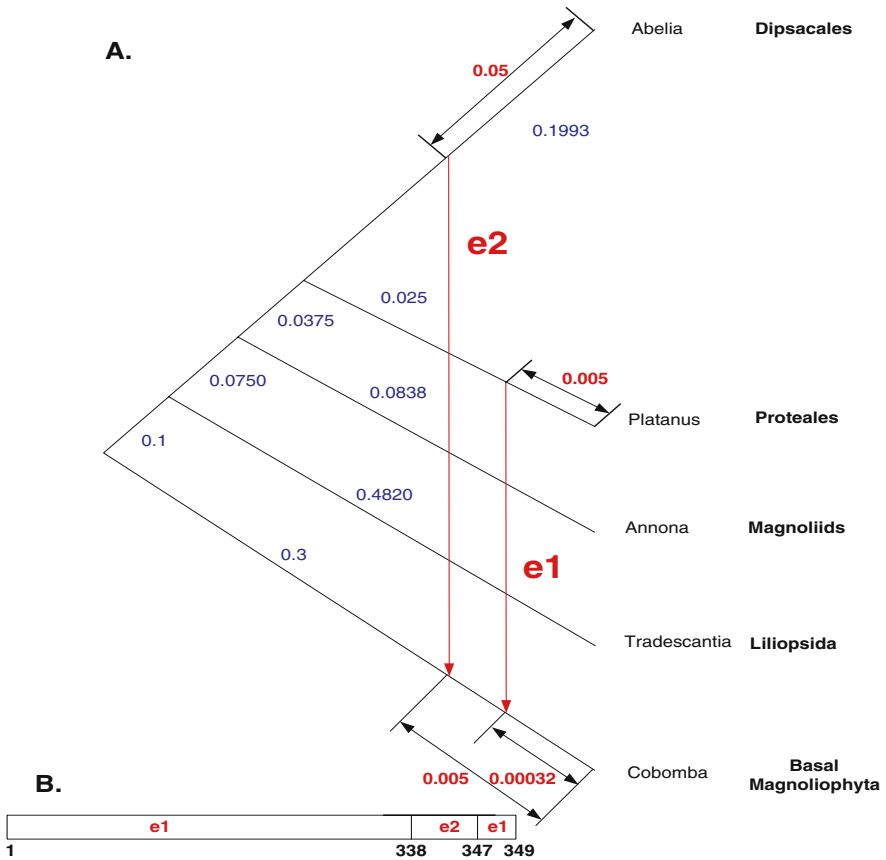


Fig. 6. Phylogenetic network of the ribosomal protein gene *rps11* of a group of 5 flowering plants. *A.* The inferred positions of the HGTs. *B* The inferred most likely path (the most likely reticulation edge in each position of the sequence).

In order to adjust the tree’s branch lengths to the evolutionary rate of our set of genes/proteins, we used one scaling factor for all the branches that was estimated together with the other network parameters (similar idea to the proportional branch lengths approach [24]). The set of the reticulation edges (without their positions along the tree branches) was inferred by the fast algorithm that is based on the MP criterion [17].

The Ribosomal Protein Gene *rps11* of Flowering Plants. We analyzed the ribosomal protein gene *rps11* of a group of 5 flowering plants which was first analyzed by Bergthorsson *et al.* [2] who suggested that this dataset underwent chimeric HGTs. This dataset consists of 5 DNA sequences. The species tree was reconstructed based on various sources, including the work of [20], the edge lengths were computed by the gene *atp1* (1, 254 nucleotides) (see figure 6).

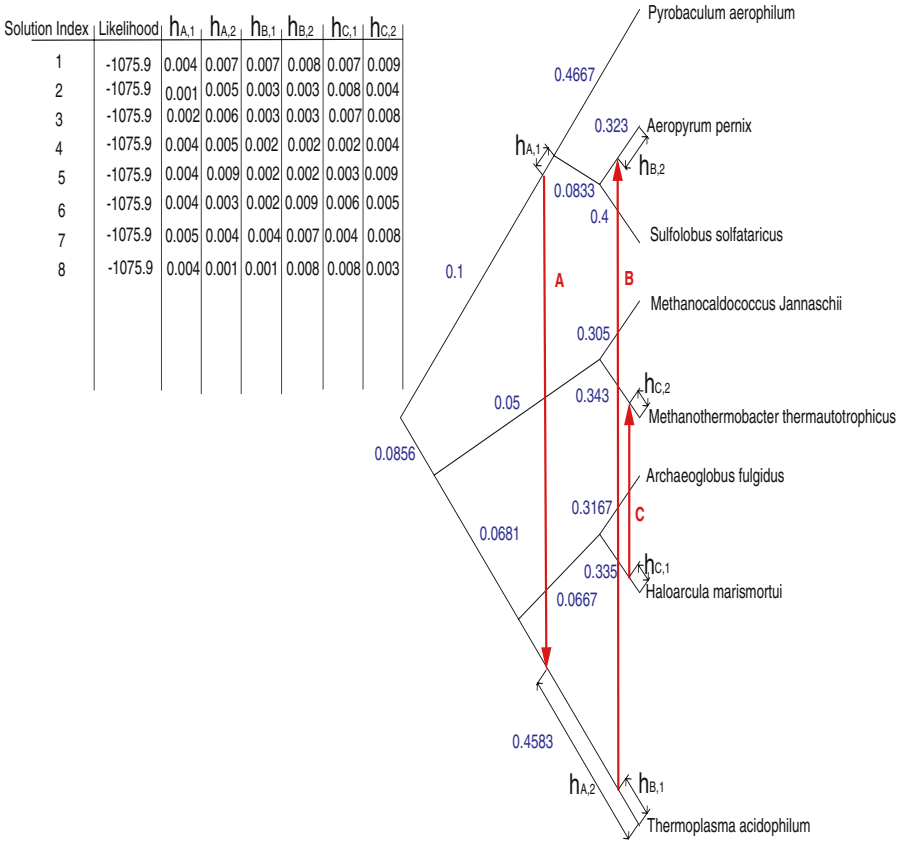


Fig. 7. Phylogenetic network of the ribosomal protein *rpl12e* of a group of 8 *Archaeal* organisms. The position of the HGTs ($h_{A,1}$, $h_{A,2}$, $h_{B,1}$, $h_{B,2}$, $h_{C,1}$, $h_{C,2}$) for each of the eight most likely solutions are plotted in the table beside.

The permutation p-value of the results is 0.02, and the plot of our EM NET-HMM supports the hypothesis of Bergthorsson *et al.* as the resulted ML paths consist of two different non-organismal trees (see figure 6). The tree that consists of one of the reticulation edges appear in most of the sites of the path but towards the end of the path the second tree (consisting of the second reticulation edge) appears (see figure 6).

Ribosomal Protein *rpl12e* of a Group of *Archaeal* Organisms. Next we analyzed the ribosomal protein *rpl12e* of a group of 8 *Archaeal* organisms, which was analyzed by Matte-Tailliez *et al.* [22]. This dataset consists of 14 aligned amino acid sequences, each of length 89 sites. We used the same organismal tree used by the authors [22]. In their work, they suggest that ribosomal protein *rpl12e* has different evolution history from the organismal evolutionary tree

(probably due to HGT events). By using MP Jin *et al.* [17] indeed found three HGTs (see figure 7), that can explain the difference between the two trees.

The inferred positions of the HGTs are depicted in figure 7. All the ML solutions have permutation p-values around 0.02 (due to almost identical log likelihoods). The distances of events B and C from the leaves are very small (see $h_{B,1}$, $h_{B,2}$, $h_{C,1}$, $h_{C,2}$ in figure 7), suggesting that these two HGT events occurred fairly recently (in terms of the evolutionary time scale).

Amelioration [15,25] is a process by which a gene that was transferred horizontally acquires features (*e.g.* GC content, the percentage of nitrogenous bases on a DNA molecule which are either guanine or cytosine) similar to its new environment. This is particularly true in recent events as this process diminishes in time. The lengths along the branches of a phylogenetic network are related to the rate of mutation and time span [21]. The existence of two paths along a tree with the same time span but different lengths suggests a variation in mutation rate. By definition, the time span between the two ends of a reticulation edge and their corresponding leaves of the phylogenetic network is identical. Interestingly, our results show $h_{B,1} < h_{B,2}$ and $h_{C,1} < h_{C,2}$ ². This fact suggests that after horizontal transfer events, genes have undergone an accelerated evolution or adaptation on the amino acid (protein) levels and not only on the nucleotides (gene) levels. This reasonable idea is new in this context.

4 Conclusions

In this work we have described a novel model for analyzing phylogenetic networks. We show that our model, along with its implementation, has advantages over other methods and it is complementary to the other methods in the field, by utilizing and extracting more information encompassed in the data. We also devised a novel test for the statistical significance of a hypothesis (network) under that model. The main strength of this model is its accuracy, however at the cost of increased complexity. The latter can be overcome by incorporating prior knowledge obtained by simple, less accurate models. We have devised an inference method for that model and have shown its performance on simulated data and subsequently applied it to real biological data that was analyzed previously by other methods. Using it, we are able to answer real biological questions such as existence of partial HGT, differences in the rate of mutation among various lineages, distribution of HGT over time and alike. On the computational side, we devised a novel algorithm, EM NET-HMM that employs an EM algorithm in the transition probability space combined with a hill climbing step in the network parameter space.

As a future work, it would be desirable to develop more efficient heuristics for optimizing our computations, and to implement our method on larger sets of organisms. By implementing the NET-HMM on large datasets we intend to

² Precisely, averaging over all ML solutions, $h_{B,1} = 0.003$, $h_{B,2} = 0.0053$, $h_{C,1} = 0.0056$, $h_{C,2} = 0.0063$.

answer the basic question of determining the length distribution of HGTs or partial HGTs.

Acknowledgment

T.T. was supported by the Edmond J. Safra Bioinformatics program at Tel Aviv University.

References

1. Addario-Berry, L., Hallett, M., Lagergren, J.: Towards identifying lateral gene transfer events. In: PSB 2003, pp. 279–290 (2003)
2. Bergthorsson, U., Adams, K., Thomason, B., Palmer, J.: Widespread horizontal transfer of mitochondrial genes in flowering plants. *Nature* 424, 197–201 (2003)
3. Birin, H., Gal-Or, Z., Elias, I., Tuller, T.: Inferring horizontal transfers in the presence of rearrangements by the minimum evolution criterion. *Bioinformatics* 24(6), 826–832 (2008)
4. Boc, A., Makarenkov, V.: New efficient algorithm for detection of horizontal gene transfer events. In: Benson, G., Page, R.D.M. (eds.) WABI 2003. LNCS (LNBI), vol. 2812, pp. 190–201. Springer, Heidelberg (2003)
5. Delwiche, C., Palmer, J.: Rampant horizontal transfer and duplication of rubisco genes in eubacteria and plastids. *Mol. Biol. Evol.* 13(6) (1996)
6. Doolittle, W.F., Boucher, Y., Nesbo, C.L., Douady, C.J., Andersson, J.O., Roger, A.J.: 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)
7. Durbin, R., Eddy, S.R., Krogh, A., Mitchison, G.: *Biological Sequence Analysis: Probabilistic Models of Proteins and Nucleic Acids*. Cambridge University Press, Cambridge (1999)
8. Paulsen, I.T., et al.: Role of mobile DNA in the evolution of Vacomycin-resistant *Enterococcus faecalis*. *Science* 299(5615), 2071–2074 (2003)
9. Felsenstein, J.: Evolutionary trees from DNA sequences: A maximum likelihood approach. *J. Mol. Evol.* 17, 368–376 (1981)
10. Hallett, M., Lagergren, J.: Efficient algorithms for lateral gene transfer problems. In: RECOMB 2001, pp. 149–156. ACM Press, New York (2001)
11. Hallett, M., Lagergren, J., Tofigh, A.: Simultaneous identification of duplications and lateral transfers. In: Proceedings of the eighth annual international conference on Research in computational molecular biology, pp. 347–356 (2004)
12. Hein, J.: Reconstructing evolution of sequences subject to recombination using parsimony. *Math. Biosci.* 98, 185–200 (1990)
13. Husmeier, D., McGuire, G.: Detecting recombination with MCMC. *Bioinformatics* 18, 345–353 (2002)
14. Huson, D.H., Bryant, D.: Application of phylogenetic networks in evolutionary studies. *Mol. Biol. Evol.* 23(2), 254–267 (2006)
15. Lawrence, J.G., Ochman, H.: Amelioration of bacterial genomes: rates of change and exchange. *J. Mol. Evol.* 44(4), 383–397 (1997)
16. Jin, G., Nakhleh, L., Snir, S., Tuller, T.: Maximum likelihood of phylogenetic networks. *Bioinformatics* 22(21), 2604–2611 (2006)

17. Jin, G., Nakhleh, L., Snir, S., Tuller, T.: Inferring phylogenetic networks by the maximum parsimony criterion: A case study. *Mol. Biol. Evol.* 24(1), 324–337 (2007)
18. Jin, G., Nakhleh, L., Snir, S., Tuller, T.: A new linear-time heuristic algorithm for computing the parsimony score of phylogenetic networks: Theoretical bounds and empirical performance. In: Măndoiu, I.I., Zelikovsky, A. (eds.) *ISBRA 2007. LNCS (LNBI)*, vol. 4463, pp. 61–72. Springer, Heidelberg (2007)
19. Jin, G., Nakhleh, L., Snir, S., Tuller, T.: Parsimony score of phylogenetic networks: Hardness results and a linear-time heuristic (submitted, 2008)
20. Judd, W.S., Olmstead, R.G.: A survey of tricolpate (eudicot) phylogenetic relationships. *Am. J. Bot.* 91, 1627–1644 (2004)
21. Jukes, T., Cantor, C.: Evolution of protein molecules. In: Munro, H.N. (ed.) *Mammalian protein metabolism*, pp. 21–132 (1969)
22. Matte-Tailliez, O., Brochier, C., Forterre, P., Philippe, H.: Archaeal phylogeny based on ribosomal proteins. *Mol. Biol. Evol.* 19(5), 631–639 (2002)
23. Nakhleh, L., Ruths, D., Wang, L.S.: RIATA-HGT: A Fast and Accurate Heuristic for Reconstructing Horizontal Gene Transfer. In: Wang, L. (ed.) *COCOON 2005. LNCS*, vol. 3595, pp. 84–93. Springer, Heidelberg (2005)
24. Pupko, T., Huchon, D., Cao, Y., Okada, N., Hasegawa, M.: Combining multiple datasets in a likelihood analysis: which models are best. *Mol. Biol. Evol.* 19(12), 2294–2307 (2002)
25. Ragan, M.A.: On surrogate methods for detecting lateral gene transfer. *FEMS Microbiol. Lett.* 201(2), 187–191 (2001)
26. Richardson, A.O., Palmer, J.D.: Horizontal gene transfer in plants. *J. Exp. Bot.* 58(1), 1–9 (2007)
27. Siepel, A., Haussler, D.: Combining phylogenetic and hidden markov models in biosequence analysis. In: *RECOMB 2003*, pp. 277–286 (2003)
28. Strimmer, K., Moulton, V.: Likelihood analysis of phylogenetic networks using directed graphical models. *Mol. Biol. Evol.* 17(6), 875–881 (2000)
29. von Haeseler, A., Churchill, G.A.: Network models for sequence evolution. *J. Mol. Evol.* 37, 77–85 (1993)