Journal of Bioinformatics and Computational Biology
Vol. 7, No. 4 (2009) 625–644

Imperial College Press
www.icpress.co.uk

THE NET-HMM APPROACH: PHYLOGENETIC NETWORK INFERENCE BY COMBINING MAXIMUM LIKELIHOOD AND HIDDEN MARKOV MODELS*

SAGI SNIR[†,§] and TAMIR TULLER[‡,¶]

[†]*Department of Evolutionary and Environmental Biology, and Institute of Evolution, University of Haifa, Israel*

[‡]*School of Computer Science, Tel Aviv University, Israel*
[§]*ssagi@research.haifa.ac.il*
[¶]*tamirtul@post.tau.ac.il*

Received 16 May 2008
Revised 5 December 2008
Accepted 6 December 2008

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, the NET-HMM, for analyzing and modeling phylogenetic networks. 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 dataset. 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 three different biological inputs.

Keywords: Phylogenetic network; Maximum Likelihood; Hidden Markov Models; horizontal gene transfer.

*A preliminary version of this work was presented at WABI08; S.S. and T.T contributed equally to this work; correspondence should be address to S.S or to T.T.

1. Introduction

Eukaryotes evolve largely through vertical lineal descent in a tree-like manner. However, in the presence of horizontal gene transfer (HGT), the right model of evolution is not a tree but is rather a *phylogenetic network*, which is a directed acyclic graph obtained by positioning a set of edges between pairs of the branches of an organismal tree to model the horizontal transfer of genetic material.[1] 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*,[2–4] but evidence of HGT (partial or complete) between Eukaryotes is also accumulating.[5,6] 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 Ref. 7) which are graphical models that capture incompatibilities in the data due to various factors, not necessarily HGT or hybrid speciation; Maximum Parsimony (MP)[1,8,9] that are based on Occam's Razor approach; distance methods, that try to fit a distance matrix to a network,[10,11] and graph theoretical approaches that try to fit a gene tree to a species tree.[12,13]

One of the most accurate and commonly used criteria for reconstructing phylogenetic trees is *Maximum Likelihood* (ML).[14] 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[15] provided a framework for evaluating likelihood on networks and subsequently[16] provided an approach to assess this likelihood. These works consider a network as an arbitrary set of splits that do not correspond to a specific biological process. Likelihood on networks has also been considered in the setting of recombination networks (see e.g. Ref. 17). 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.[1] 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 Siepsel and Haussler[18] 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

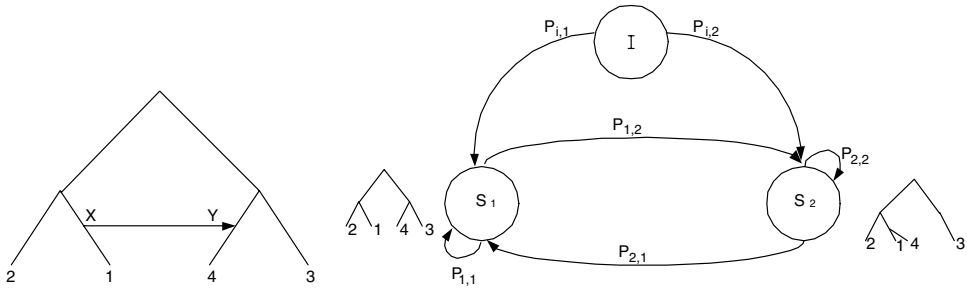


Fig. 1. Simple example of the NET-HMM model. (a) A phylogenetic network, with a single HGT from X to Y. (b) The HMM for the network: I denotes the initial state; all the other states are related to the networks' trees. The state S_1 is related to the underlining organismal tree. The state S_2 is related to the tree that realizes 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 another tree.

in Ref. 1 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 Fig. 1), i.e. the emission probability in each state is according to its corresponding tree.

The model is supported by biological facts that (i) adjacent sites are not independent with respect to HGT, and (ii) events of HGT are relatively rare. In the view of the NET-HMM model, events such as chimeric HGT[19] 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[9,10] cannot accomplish.

By applying the NET-HMM on synthetical data, we show very significant improvements over the i.i.d. model[1] in the ability to locate chimeric events and locations along tree edges. Results on biological data point out significant biological phenomena such as amelioration,[20,21] chimeric HGT events, and occurrence of very recent HGT events. Applying the NET-HMM to data that were analyzed in the past suggests 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 such that 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 $\mathbf{q} : E \to \mathbb{R}$ 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.[22] 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 ith site, S_i, is the set of states at the ith position for every sequence in S.[a] 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[14]:

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

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, such that 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 Refs. 23 and 24). This results in having no cycles in which tree edges are traversed along their directionality and reticulation

[a]It can be viewed as the ith column when the sequences are aligned.

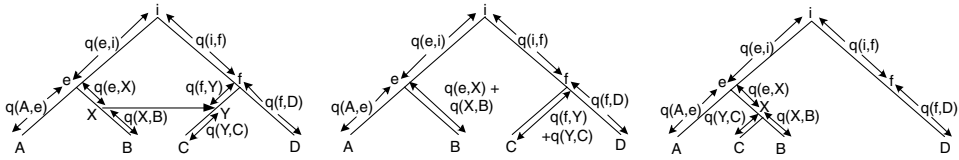


Fig. 2. A simple example of a phylogenetic network under the likelihood setting, and the set of induced trees. (a) A phylogenetic network with one reticulation edge; each edges length denotes the expected number of subsections along the edge. (b) One of the network's trees that does not include the reticulation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

edges in either direction [see Fig. 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 Fig. 2).

2.2. *From a network to an 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 such that the 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[1]: $L(S_i|N) = \sum_{t \in T(N)} P(E_t) \cdot L(S_i|t)$. 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 which 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 ith site to a state of the HMM is that this site evolves on the tree $\Upsilon(h) \in T(N)$ (i.e. the ith column was emitted by the tree T).

Let $p(h_{i-1} \to h_i)$ 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

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

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 [Eq. (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 Ref. 1).

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 Fig. 2).
(2) Temporal constraints on the reticulation edges decrease the number of valid networks [see Fig. 3(a)].
(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 Fig. 3(b)]. More details about techniques to reduce the complexity of the model appear in Sec. 2 of the Supplementary Material.

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

(1) **The tiny version.** Input: A set of sequences associated with the tree leaves, 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: A set of sequences associated with the tree leaves, the network topology. Output: The ML network edge probabilities and ML transition probabilities of the induced HMM.
(3) **The large version.** Input: A set of sequences associated with the tree leaves, the initial organismal tree. Output: The ML NET-HMM (complete network + HMM; i.e. the network topology along with ML network edges probabilities and ML transition probabilities in the induced HMM).

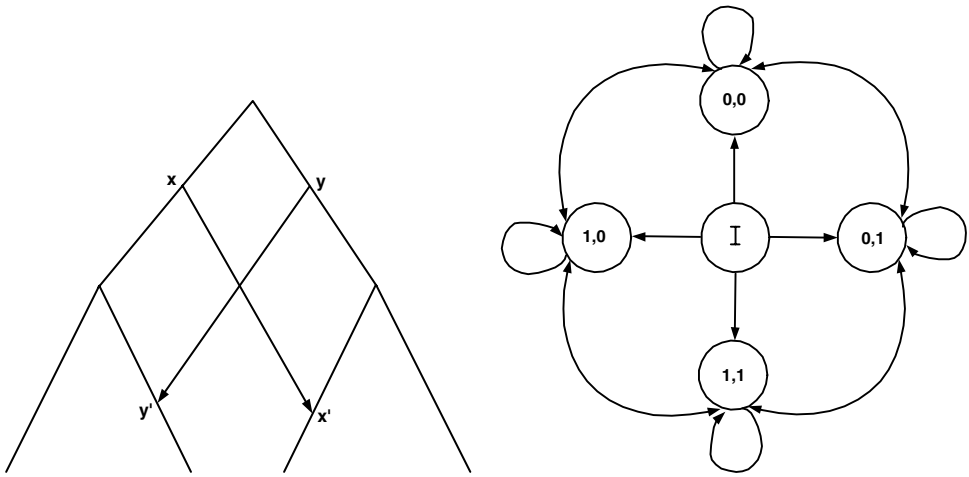


Fig. 3. Constraints on the structure of the NET-HMM. (a) Example of a network that does not satisfy the temporal constraints. By the tree topology, x occurs before y' and y occurs before x'. By the reticulation edges, x and x' occur at the same time and y and y' occur at the same time. Thus, y' occurs before x' and x' occurs before y' — a contradiction. (b) A simple example of the implementation of the assumption that HGT's are rare events. Suppose the phylogenetic network includes two reticulation 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.

The specific version of the problem that should be employed depends 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.

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 ith 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

$$\sum_{h'' \in \Sigma_H} L(S_1^{i-1}|h'', M) \cdot p(h'' \to h') \cdot L(S_i|\Upsilon(h'), \mathbf{q}). \tag{3}$$

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 Eq. (2) by hill climbing.

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

Equation (3) is solved by the dynamic programming forward and backward algorithms.[25] 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 Fig. 4), for estimating the edge lengths and edge probabilities of the network as well as the state transition probabilities in the HMM (the EM algorithm of Baum–Welch is described in the Supplementary Material).

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 such that 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 Refs. 1 and 9). We however used a more rigorous approach as described in the next subsection.

2.4. *Network significance*

Recall that the improvement of the NET-HMM model over the i.i.d. (see Sec. 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' with respect to 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 permutation S_0 of the sites of the input, let $L(S_0|M')$ be the likelihood of M' with respect to S_0 (i.e. the solution of the tiny problem). For a given big enough set of such permutations, we obtain a probability distribution, of the likelihood of M'. Given that distribution, we can compute the *empirical p-value* of M' with respect to 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 expect that the likelihood of a model M' given the data S_0 will be higher than the likelihood of M' given a permuted version of the input data, S_0. Thus, as we demonstrate in this work, this approach can be useful in determining a stopping criterion. We stopped the algorithm of iteratively adding HGT edges when either the likelihood did not increase or this empirical p-value was not-significant.

Note that a stopping criterion which is based on Minimum Description Length (MDL; see for example, Ref. 26) does not give good results in practice. It appeared that the MDL needs some scaling that is specific to each dataset (similar problem was observed in Ref. 1). Thus, we decide not to use it (see the Supplementary Material for more details).

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.

Each of the synthetical phylogenetic networks included eight taxa and two reticulation edges. The tree edge lengths were sampled from the uniform distribution $U[0, 0.25]$. Thus, each of these networks included four trees (see Fig. 5). To simulate genomic sequences where each part of the sequence was evolved along different tree of the network, we sampled segments (consecutive sets of sites) from each of the networks' trees and concatenated them into a basic concatenation block (see Fig. 5). The experiments were conducted for various HGT segment lengths and for various number of replications of the basic concatenation block (see Fig. 5).

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) that occurred in the network. This is equivalent to finding the correct path between the HMM states (trees). 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 results are shown in Fig. 6.

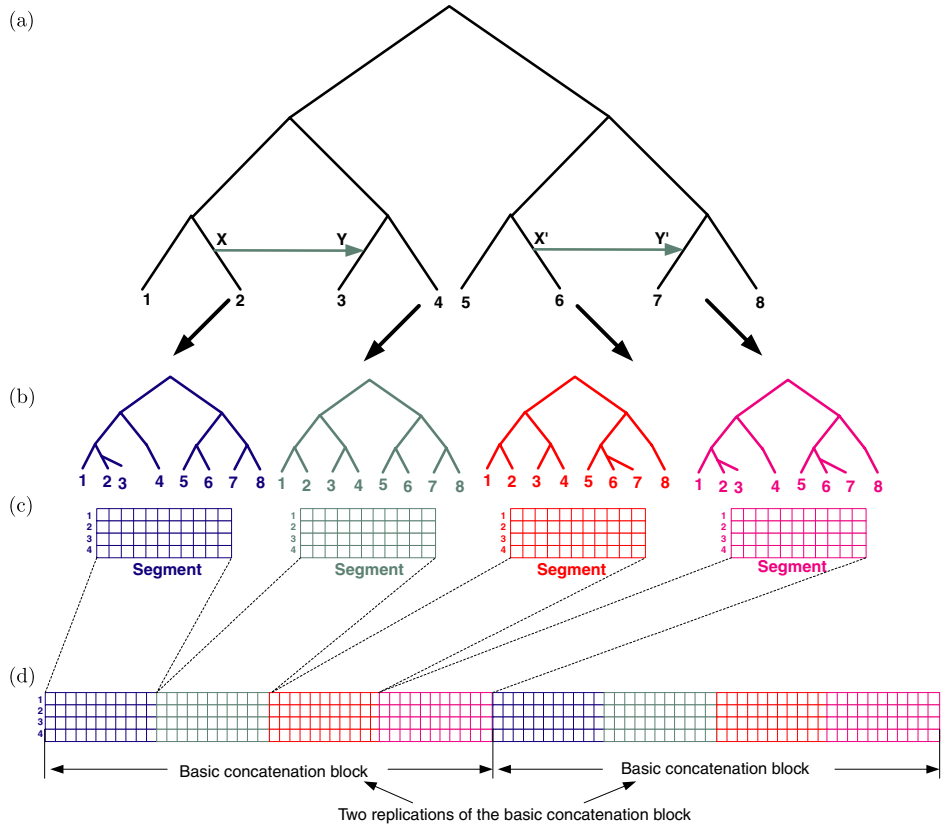
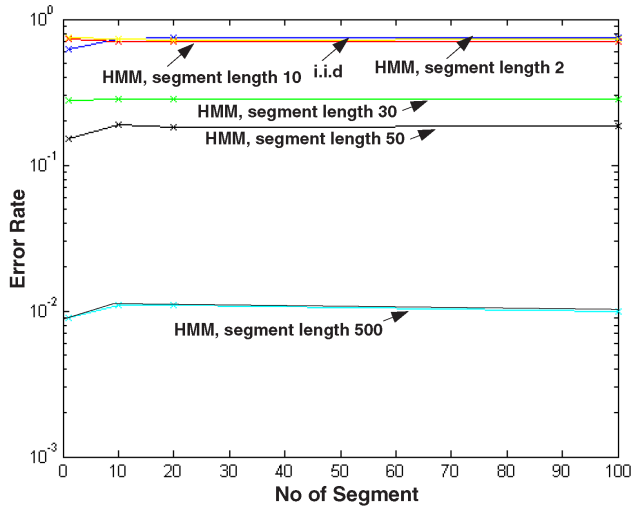
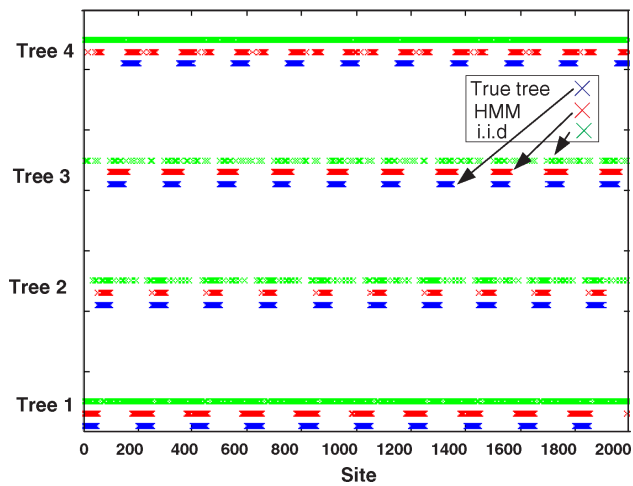


Fig. 5. Illustration of the simulation. (a) The input includes phylogenetic networks, each with eight taxa and two reticulation edges. (b) Each network includes four trees. (c) Segments were sampled from each of the trees. (d) These segments were concatenated; each input to our method included the concatenation of the corresponding segments and the topology of the corresponding network.

In Fig. 6(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 6(b) and 6(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.

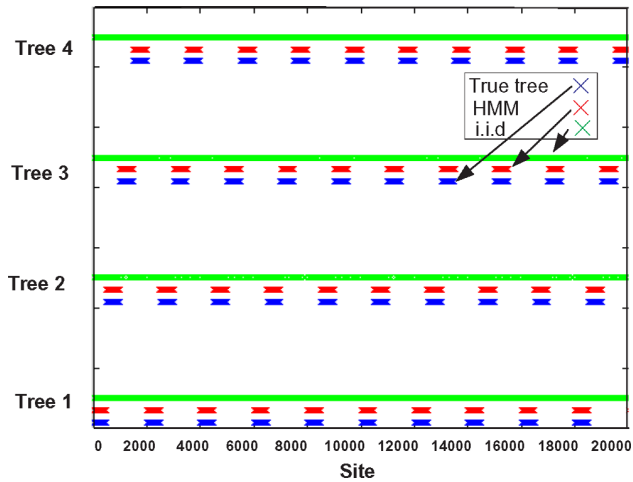


(a)



(b)

Fig. 6. The performances of the NET-HMM on synthetic inputs (a network with eight leaves and two reticulation edges). The x-axis is the position along the input sequences (the index of a site). The y-axis is the network's tree that corresponds to each site; as there are two reticulation edges the network induces four trees (named Tree 1, . . . , Tree 4). For each site, we marked the true tree corresponding to this site (blue x near the corresponding tree), the tree that the i.i.d model inferred for this site (green x near the corresponding tree), and the tree that the NET-HMM inferred for this site (red x near the corresponding tree). (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 versus the i.i.d model. As can be seen, the trees inferred by the NET-HMM are very close to the true trees; the trees inferred by the i.i.d. model are scattered over the y-axis. (c) Segment length 500: The inferred tree in each site by NET-HMM versus the i.i.d model. This figure is even clearer than (b): the trees inferred by the NET-HMM are very close to the true trees; the trees inferred by the i.i.d. model are scattered over the y-axis.



(c)

Fig. 6. (*Continued*)

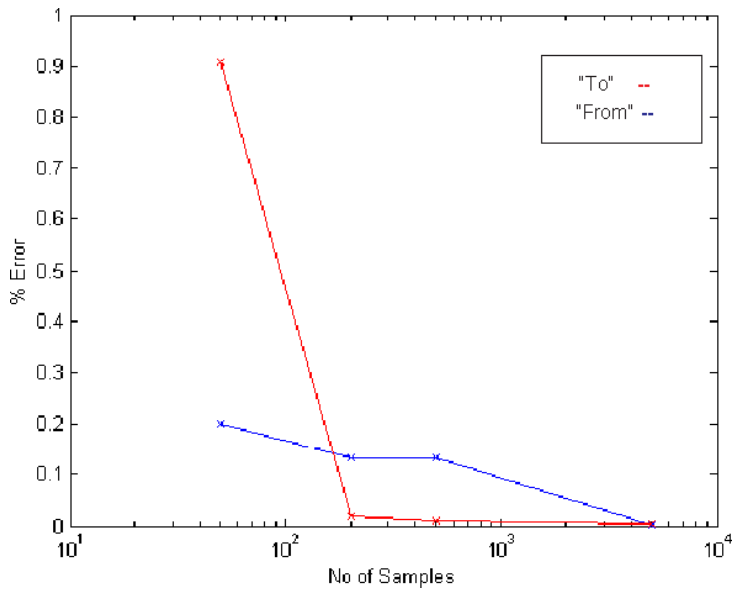


Fig. 7. Convergence of location of reticulation events (averaged over 20 networks with eight leaves and two reticulation edges). The error rate is measured as the distance from the true location along the tree edge divided by the edge length.

In our second test, we examined the ability of the NET-HMM method to infer the exact location of the HGT events (modeled by the reticulation edges) on the tree edges. The results are presented in Fig. 7.

We found a difference in the convergence rate between the "from" and "to" edges (the tree edges from which the HGT edges emanated and those to which

they are connected respectively). While the "from" error is always very low (below 20%) the "to" error starts very high at sequence length 50 but diminishes very sharply towards sequence length 200. In general, for realistic HGT lengths (200), both error rates are at a very satisfactory level (below 12%).

Finally, 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.

Running times. As mentioned before, the running times of the algorithm are reasonable for analyzing 7–10 sequences with up to four edges. The size of the network (No. of trees) grows exponentially in the number of HGT edges k and is bounded by 2^k. The Forward/Backward algorithm runs in time quadratic in the size of the network and linear in the sequence lengths ℓ, that is $O(\ell 2^{2k})$. The EM NET-HMM runs the Forward/Backward step until convergence is reached for every new HGT edge added. In our experiments this was approximately $12 \cdot \ell 2^{2k}$msec (approximately 1.5 minutes for an edge in a network with two reticulation edges and sequence length 500). To avoid local ML points, we repeated this process ten times, yielding a constant factor of 120. As a tree over n taxa has at most $4n^2$ pairs of tree edges ("from" and "to") the possible location grows quadratically with n. Therefore we obtained approximate typical running times for the big problem of about $k \cdot n^2 \cdot 120 \cdot \ell \cdot 2^{2k}$ msec for adding k reticulation edges to a tree over n sequences of length ℓ each. Indeed running the big problem for inserting two reticulation edges on our typical dataset of eight sequences of length 500, took about a complete day.

3.2. *Biological inputs: the small problem*

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.

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[27]). 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.[9]

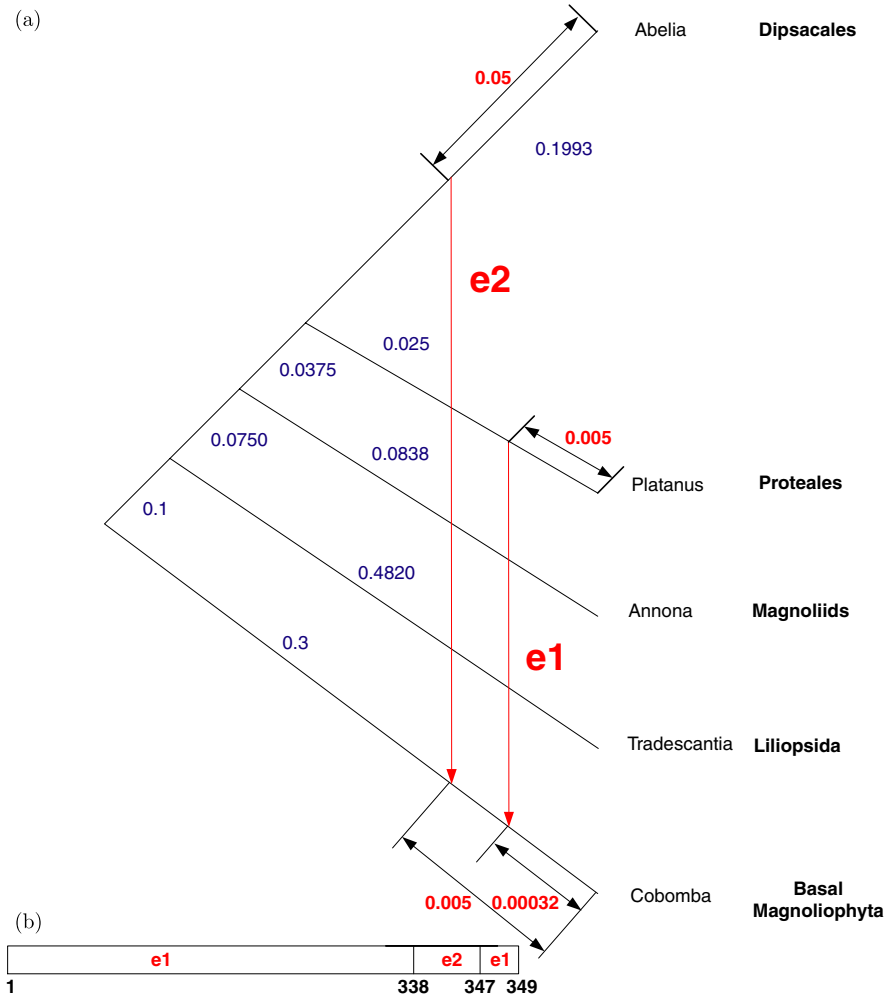


Fig. 8. Phylogenetic network of the ribosomal protein gene *rps11* of a group of five flowering plants. (a) The inferred positions of the HGTs. (b) The inferred most likely path (the most likely reticulation edge in each position of the sequence).

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.*[5] who suggested that this dataset underwent chimeric HGTs. This dataset consists of five DNA sequences. The species tree was reconstructed based on various sources, including the work of,[28] the edge lengths were computed by the gene *atp*1 (1254 nucleotides) (see Fig. 8).

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 Fig. 8). The tree that consists of

one of the reticulation edges appears in most of the site of the path, but towards the end of the path the second tree (consisting of the second reticulation edge) appears (see Fig. 8).

Ribosomal protein *rpl12e* of a group of *Archaeal* organisms. Next we analyzed the ribosomal protein *rpl12e* of a group of eight *Archaeal* organisms, which was analyzed by Matte-Tailliez *et al.*[19] This dataset consists of 14 aligned amino acid sequences, each of length 89 sites. We used the same organismal tree used by the authors.[19] 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.*[9] indeed found three HGTs (see Fig. 9) that can explain the difference between the two trees.

The inferred positions of the HGTs are depicted in Fig. 9. 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 Fig. 9), suggesting that these two HGT events occurred fairly recently (in terms of the evolutionary time scale).

Amelioration[20,21] 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

Solution Index	Likelihood	$h_{A,1}$	$h_{A,2}$	$h_{B,1}$	$h_{B,2}$	$h_{C,1}$	$h_{C,2}$
1	−1075.9	0.004	0.007	0.007	0.008	0.007	0.009
2	−1075.9	0.001	0.005	0.003	0.003	0.008	0.004
3	−1075.9	0.002	0.006	0.003	0.003	0.007	0.008
4	−1075.9	0.004	0.005	0.002	0.002	0.002	0.004
5	−1075.9	0.004	0.009	0.002	0.002	0.003	0.009
6	−1075.9	0.004	0.003	0.002	0.009	0.006	0.005
7	−1075.9	0.005	0.004	0.004	0.007	0.004	0.008
8	−1075.9	0.004	0.001	0.001	0.008	0.008	0.003

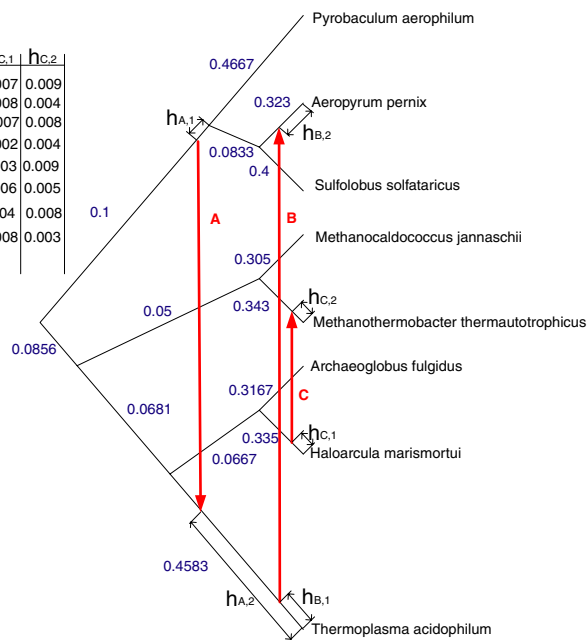


Fig. 9. Phylogenetic network of the ribosomal protein *rpl12e* of a group of eight *Archaeal* organisms. The position of the HGTs ($h_{A,1}, h_{A,2}, h_{B,1}, h_{B,2}, h_{C,1}, h_{C,2}$) for each of the eight most likely solutions are plotted in the table beside.

is particularly true for 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.[22] The existence of two paths along a tree with the same time span but with 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 that $h_{B,1} < h_{B,2}$ and $h_{C,1} < h_{C,2}$.[b] 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.

3.3. *Biological inputs: the big problem*

As mentioned, we think that the most appropriate use of our method is via the small problem. However, in this section we demonstrate solving the big problem on two biological inputs, 8 *Archaea*, and 8 *Wolbachia*.

Ribosomal protein *rpl12e* of a group of *Archaeal* organisms. Here we solve the big problem on the *Archaea* (8 taxa) dataset that was described before. The initial log likelihood (of the organismal tree) was -1107.15, the first reticulation edge (edge A in Fig. 9) improved the log likelihood to -1089.2 (p-value 0.04), the second reticulation edge (edge B in Fig. 9) improved the log likelihood to -1076.7 (p-value 0.03); the improvement of the third reticulation edge (edge C in Fig. 9) was minor (the result log likelihood was -1075.9, p-value 0.032). These results suggest that only the first two HGTs are significant. This finding enforces the claim that the third additional HGT event found by MP in Ref. 9 is potentially spurious.

Gene *gltA* of a group of *Wolbachia*. Baldo *et al.*[29] analyzed this gene (and three other housekeeping genes) on a dataset of eight *Wolbachia*, bacterial symbionts of arthropods. They suggest that the gene underwent horizontal transfer(s) of genetic material via homologous recombination, thus it has a multiple origin.

The DNA sequences of the four genes were downloaded from NCBI (http://ncbi.nih.gov) (accession numbers appear in Ref. 29); the organismal tree was based on the concatenated alignment of these four *Wolbachia* genes.[29]

The analysis of Baldo *et al.*[29] was based on comparing the phylogenetic trees of these different genes ($gltA, dnaA, groEL, ftsZ$), analysis of genetic variation,[30] and usage of four programs for detecting recombination in aligned sequences.[31–33] Our results support the results of the analysis of Baldo *et al.* We found two reticulation edges that significantly improve the likelihood of the model (see Fig. 10); the first edge is from *A. eponina* to *P. sialia* while the second is from *C. cautella* to *D. simulans*.

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

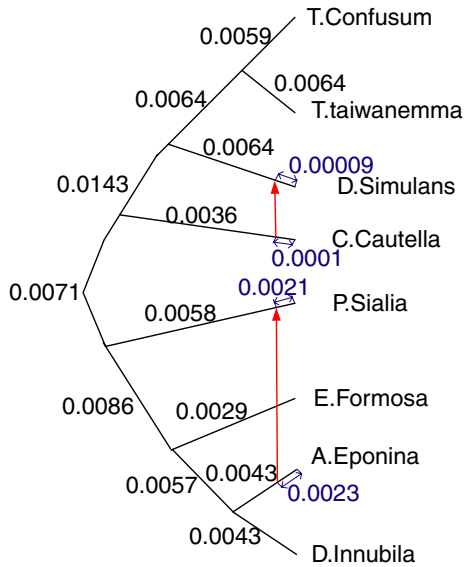


Fig. 10. Two reticulation edges that were found for the *Wolbachia* dataset. The first reticulation edge (from *A. eponina* to *P. sialia*) improved the log likelihood from −4230 to −3955 (p-value = 0.05). The second reticulation edge improved the log likelihood to −3830 (p-value = 0.05), the third reticulation edge did not give a significant improvement (log likelihood to −3830.8, p-value = 0.05).

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

Supplementary Material

Supplementary material of the work including more plots of the algorithm is available at *http://www.cs.tau.ac.il/ ∼ tamirtul/NETHMM.html*

Acknowledgment

T.T. was supported by the Edmond J. Safra Bioinformatics Program at Tel Aviv University and the Yeshaya Horowitz Association through the Center for Complexity Science.

References

1. Jin G, Nakhleh L, Snir S, Tuller T, Maximum likelihood of phylogenetic networks, *Bioinformatics* **22**(21):2604–2611, 2006.
2. Doolittle WF, Boucher Y, Nesbo CL, Douady CJ, Andersson JO, Roger AJ, 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.
3. Paulsen IT *et al.*, Role of mobile DNA in the evolution of Vancomycin-resistant *Enterococcus faecalis*, *Science* **299**(5615):2071–2074, 2003.
4. Delwiche C, Palmer J, Rampant horizontal transfer and duplication of rubisco genes in eubacteria and plastids, *Mol Biol Evol* **13**(6), 1996.
5. Bergthorsson U, Adams K, Thomason B, Palmer J, Widespread horizontal transfer of mitochondrial genes in flowering plants, *Nature* **424**:197–201, 2003.
6. Richardson AO, Palmer JD, Horizontal gene transfer in plants, *J Exp Bot* **58**(1):1–9, 2007.
7. Huson DH, Bryant D, Application of phylogenetic networks in evolutionary studies, *Mol Biol Evol* **23**(2):254–267, 2006.
8. Hein J, Reconstructing evolution of sequences subject to recombination using parsimony, *Math Biosci* **98**:185–200, 1990.
9. 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.
10. Boc A, Makarenkov V, New efficient algorithm for detection of horizontal gene transfer events, *WABI*, pp. 190–201, 2003.
11. 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.
12. Addario-Berry L, Hallett M, Lagergren J, Towards identifying lateral gene transfer events, *PSB03*, pp. 279–290, 2003.
13. Hallett M, Lagergren J, Efficient algorithms for lateral gene transfer problems, *RECOMB01*, pp. 149–156, ACM Press, New York, 2001.
14. Felsenstein J, Evolutionary trees from DNA sequences: A maximum likelihood approach, *J Mol Evol* **17**:368–376, 1981.
15. von Haeseler A, Churchill GA, Network models for sequence evolution, *J Mol Evol* **37**:77–85, 1993.
16. Strimmer K, Moulton V, Likelihood analysis of phylogenetic networks using directed graphical models, *Mol Biol Evol* **17**(6):875–881, 2000.
17. Husmeier D, McGuire G, Detecting recombination with MCMC, *Bioinformatics* **18**:345–353, 2002.

18. Siepel A, Haussler D, Combining phylogenetic and Hidden Markov Models in biosequence analysis, *RECOMB03*, pp. 277–286, 2003.
19. Matte-Tailliez O, Brochier C, Forterre P, Philippe H, Archaeal phylogeny based on ribosomal proteins, *Mol Biol Evol* **19**(5):631–639, 2002.
20. Lawrence JG, Ochman H, Amelioration of bacterial genomes: Rates of change and exchange, *J Mol Evol* **44**(4):383–397, 1997.
21. Ragan MA, On surrogate methods for detecting lateral gene transfer, *FEMS Microbiol Lett* **201**(2):187–191, 2001.
22. Jukes T, Cantor C, Evolution of protein molecules, in Munro HN, Allison JB, (eds.), *Mammalian Protein Metabolism*, Academic Press, Academic Press, pp. 21–132, 1969.
23. Jin G, Nakhleh L, Snir S, Tuller T, Parsimony score of phylogenetic networks: Hardness results and a linear-time heuristic, *IEEE/ACM Trans Comput Biol Bioinform*, Accepted, October 2008.
24. 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, *ISBRA*, pp. 61–72, 2007.
25. Durbin R, Eddy SR, Krogh A, Mitchison G, *Biological Sequence Analysis: Probabilistic Models of Proteins and Nucleic Acids*, Cambridge University Press, 1999.
26. Rissanen J, Modeling by shortest data description, *Automatica* **14**:465–471, 1978.
27. 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.
28. Judd WS, Olmstead RG, A survey of tricolpate (eudicot) phylogenetic relationships, *Am J Bot* **91**:1627–1644, 2004.
29. Baldo L, Bordenstein S, Wernegreen JJ, Werren JH, Widespread recombination throughout *Wolbachia* genomes, *Mol Biol Evol* **23**:437–449, 2006.
30. Rozas J, Sanchez-DelBarrio JC, Messeguer X, Rozas R, Dnasp, DNA polymorphism analyses by the coalescent and other methods, *Bioinformatics* **19**:2496–2497, 2003.
31. Martin DP, Williamson C, Posada D, Rdp2: Recombination detection and analysis from sequence alignments, *Bioinformatics* **21**:260–262, 2005.
32. Martin D, Rybicki E, Rdp: Detection of recombination amongst aligned sequences, *Bioinformatics* **16**:562–563, 2000.
33. Sawyer S, Statistical tests for detecting gene conversion, *Mol Biol Evol* **6**:526–538, 1989.

Sagi Snir received his B.A. degree from Bar Ilan University at Israel, majoring in Computer Science and Economics. He received the M.S. and Ph.D. degrees in Computer Science from the Technion, Israel. He spent two years as a postdoctoral researcher in the computer science and math departments at the University of California at Berkeley, USA. He is now at the Department of Evolutionary and Environmental Biology, and the Institute of Evolution, University of Haifa, Israel. Before working on the Ph.D. degree, he worked in various information technologies companies, including IBM Haifa Research Lab. His main research interests include computational biology and, in particular, phylogenetics.

Tamir Tuller is a postdoctoral fellow in the School of Computer Science and the Department of Molecular Microbiology and Biotechnology at Tel-Aviv University, Israel. He is a fellow of the Edmond J. Safra Bioinformatics Program at Tel-Aviv University and the Yeshaya Horowitz Association through the Center for Complexity Science. Tamir received his B.S., and M.S. degrees in Electrical Engineering from Tel-Aviv University and the Technion — Israel Institute of Technology, respectively; and received the Ph.D. degree in Computer Science from Tel-Aviv University in 2006. His research interests fall in the general areas of computational biology, systems biology, and bioinformatics; in particular, he works on computational phylogenetics, analysis of co-evolution, as well as autoimmunity and bioinformatics of diseases.