# Likelihood-based Inference of Phylogenetic Networks from Sequence Data by PhyloDAG

Quan Nguyen and Teemu Roos

Helsinki Institute for Information Technology HIIT
Department of Computer Science
University of Helsinki, PO Box 68, FI-00014, Finland
{quan.nguyen,teemu.roos}@cs.helsinki.fi

**Abstract.** Processes such as hybridization, horizontal gene transfer, and recombination result in reticulation which can be modeled by phylogenetic networks. Earlier likelihood-based methods for inferring phylogenetic networks from sequence data have been encumbered by the computational challenges related to likelihood evaluations. Consequently, they have required that the possible network hypotheses be given explicitly or implicitly in terms of a backbone tree to which reticulation edges are added. To achieve speed required for unrestricted network search instead of only adding reticulation edges to an initial tree structure, we employ several fast approximate inference techniques. Preliminary numerical and real data experiments demonstrate that the proposed method, PhyloDAG, is able to learn accurate phylogenetic networks based on limited amounts of data using moderate amounts of computational resources.

**Keywords:** phylogenetic networks, likelihood-based inference, phylogenetics, probabilistic graphical models

## 1 Introduction

Phylogenetic trees are widely used for modeling the evolution of a group of organisms. However, trees are not able to represent reticulation events due to processes such as hybridization, horizontal gene transfer, and recombination. If reticulation is thought to be present, a phylogenetic network is a more useful model. For this reason, researchers in quantititive biology have been interested in representing evolutionary processes using network models since as early as the 1970s [20]. Even though various computational techniques have been proposed to deal with the challenges caused by network-like models, inferring the network structure from data remains a problem.

We propose a combination of solutions for speeding up the required computations in a likelyhood-based framework. These include a stochastic expectation-maximization (EM) algorithm for dealing with unobserved ancestral sequences. As a subroutine of the EM algorithm, we apply an approximate inference method known as loopy belief propagation [16], which provides dramatic computational

savings when computing the required sampling distributions while avoiding any unwarranted independence assumptions (see e.g., [6]).

We describe a stand-alone method, which we call PhyloDAG,[1] which can learn phylogenetic networks from data. The present implementation assumes a generic mixture model of the reticulation process but the model can be extended to handle more specific kinds of processes as well. Preliminary numerical and real world experiments demonstrate the utility of the method. For an application of PhyloDAG to the analysis of non-biological data, see [23].

The rest of the paper is organized as follows. In Sec. 2, we review some of the relevant prior work on likelihood-based phylogenetic networks. In Sec. 3, we describe our model in detail. We introduce the PhyloDAG method in Sec. 4, and present experimental results in Sec. 5. A summary and pointers for future work are given in Sec. 6.

## 2 Related Work

Likelihood-based inference has become a popular approach in phylogenetics since it was first proposed by Felsenstein [5]. Likelihood-based methods are widely considered to be the state-of-the-art in molecular phylogenetics [4, 26].

The first framework for likelihood-based inference of phylogenetic networks was proposed by Haeseler and Churchill [8]. Based on their work, Strimmer and Moulton [21] proposed to use directed graphical models, or Bayesian networks, as a represention of explicit likelihood-based phylogenetic networks. Their framework was first applied to split networks, but it can be easily applied to evolutionary networks [22]. However, networks pose major computational challenges for likelihood-based inference. Computations involving unobserved ancestral sequences are in general intractable. The solution applied in [21] is to approximate the likelihood by method similar to Gibbs sampling.

Strimmer et al. [22] model reticulation events by introducing a random variable that indicates which one of the possible ancestral taxa is active and using the same mechanism as in tree-structured models as if the active taxon were the only immediate ancestor. The random choice of the ancestor taxon is repeated independently at each site according to fixed but unknown weight parameters. The authors referred to this as the mixture model. In this work, we adopt the mixture model and develop novel efficient algorithms that can be used for inferring the network structure and parameters from data.

Jin et al. [10] point out the importance of allowing different evolutionary mechanisms for different genomic sites. However, despite their emphasis on the differences between their approach and that of Strimmer et al., the existence of a separate edge length parameter for each site, which significantly increases the model complexity but simplifies the computations, turns out to be the distinctive feature of their model. In follow-up work, Park and Nakleh [15] consider given

[1] The implementation is available for download at `http://phylomemetic.wordpress.com/2015/04/17/phylodag/` .

genomic regions inside which a fixed ancestor taxon and edge length value is used.

There are also other sophisticated ways to relax the mixture model assumption. Husmeier and Wright [9] and Webb et al. [25] assume each site to be generated from an unknown phylogenetic tree which is a hidden state in a hidden Markov model (HMM). Transitions between the states of the HMM constitute breakpoints from one phylogenetic structure to another. This approach is likely to be more realistic under recombination scenarios, but it is very computationally expensive since it introduces complex dependencies between the sites and the state space of the HMM grows exponentially in the number of taxa.

In all of the aforementioned work, due to the said computational challenges, network search is either restricted to a small set of possible networks given explicitly by the user or more implicitly to networks obtained by adding reticulation edges to a fixed backbone tree structure obtained by standard tree methods such as MrBayes [17]. A key assumption behind the use of a backbone tree is that even when the actual phylogenetic process involves reticulation events, a tree structure estimated from the data comprises a part of the true network that represents the phylogenetic history. If this is the case, the true network can be obtained by adding reticulation edges. Unfortunately, in our experience this assumption is unlikely to hold in practice. In Sec. 5.3 we demonstrate simple cases where a violation of the assumption leads to suboptimal outcomes.

Apart from horizontal gene transfer and other processes discussed above, deep coalescence arising from incomplete lineage sorting is another source of incompatibility of gene trees for individual sites or genes of a given same set of taxa, see e.g., [13]. Since deep coalescence tends to occur even when the organisms' evolution is completely tree-like, it is usually not considered to be a type of reticulation. The models used to handle deep coalescence are also somewhat distinct from those used to handle reticulation. Recently, there have been several attempts to incorporate reticulation into models for deep coalescence [12, 27].

## 3 Likelihood-based Inference in Phylogenetic Trees and Networks

We adopt the standard likelihood-based framework in phylogenetics and let each node (either leaf or internal) of a phylogenetic tree correspond to a taxon. Leaf nodes are assumed to be extant taxa whose genomic sequences are observed. In this work we focus on DNA sequences although for example protein sequences can in principle be handled in the same fashion.

We denote the probability that a DNA sequence associated to node $X_i$ in a phylogenetic tree evolves from the sequence in its immediate ancestor, called its *parent*, $\mathrm{Pa}_i$ in time proportional to branch length $\tau_i$ by $P_{\tau_i}(X_i \mid \mathrm{Pa}_i)$. These local probabilities are specified explicitly by a sequence evolution model such as the Jukes-Cantor (JC) model [11] as a function of $\tau_i$. In the following, we denote random variables and sequences like $X_i$ by upper case letters and their values, such as $x_i$, by lower case letters.

The above kind of probabilistic model describes the following evolution scenario. The nucleotide sequence at the root $X_r$ is drawn independently from a stationary distribution $\pi$ obtained as the limit $\pi(X) = \lim_{\tau\to\infty} P_\tau(X \mid y)$ for any sequence $y$. The sequence evolves independently along the edges of the tree. Assuming a fully observed tree $T$ with $p$ nodes (taxa), the likelihood of a single site at all taxa is factorized as

$$P_{(T,\tau)}(X_1 = x_1, ..., X_p = x_p) = \pi(x_r) \prod_{i \neq r} P_{\tau_i}(x_i \mid pa_i), \tag{1}$$

where $pa_i$ denotes the nucleotide at the site in question in the parent of taxon $X_i$ in tree $T$. However, since we assume that only the sequences in the leaf nodes are observed, the internal nodes, including the root node, represent ancestral taxa whose biological sequences are unavailable, and hence they become latent (unobserved) variables in the model.

Following and extending the convention familiar from phylogenetic trees, we assume that any node in a phylogenetic network is classified in one of three categories based on the number of its parents. First, the unique root node has no parents and two children (immediate descendants). Second, *tree nodes* have a single parent and either zero or two children. For both of these classes of nodes, the evolutionary model coincides with the model commonly used for likelihood-based phylogenetic trees. The third class of nodes are the *reticulation nodes* which have two parents and either zero or two children. For a given nucleotide $x_i \in \{A, C, G, T\}$ in reticulation node $X_i$, we have the conditional probability given its parents' states $pa_i = (y_i, z_i)$ as the weighted sum of its conditional probability given a single parent:

$$P_{(w_i,\tau_i)}(x_i \mid pa_i) = w_i P_{\eta_i}(x_i \mid y_i) + (1 - w_i) P_{\zeta_i}(x_i \mid z_i), \tag{2}$$

where the probabilities on the right side of the equation are the same as in the case of tree models, and the weight parameter $w_i$ as well as the edge length parameters $\tau_i = (\eta_i, \zeta_i)$ are parameters whose values need to be given in order to make the model fully specified. Plugging the above terms in the factorization (1) provides a complete probability model for reticulate evolution.

The model in Eq. (2) is the mixture model of Strimmer and Moulton [21]. If genomic regions that follow a fixed ancestry are given like in [15], they can be incorporated in the model by treating sites within a given region as a sample of data from the same source. In this work, we focus on the case where the sites are independent.

From a computational point of view, most of the complications arise from the fact that the observed-data likelihood involves a summation over the possible values of the latent variables. In tree topologies, well-known techniques exist for carrying out the summation in linear time with respect to the size of the tree [5]. These techniques are known in probabilistic graphical models more generally as *variable elimination*. Felsenstein [5] uses the expectation–maximization (EM) algorithm [3] to estimate branch length parameters in tree-structured models. In the following, we introduce methods for approximating the computations in the case where the phylogenetic hypothesis involves reticulation nodes.

## 4 The PhyloDAG Method

We propose an efficient method for likelihood-based inference of phylogenetic networks. The key novelties of the PhyloDAG method include a stochastic EM algorithm for learning the structure and parameters of the network as well as a fast loopy belief propagation (LBP) algorithm which is used to accelerate the required computations involving the latent variables in the model.

The outer loop of the algorithm is a stochastic structural EM (SSEM) algorithm. Similar to regular EM, SSEM repeats iterations consisting of an expectation (E) step followed by a maximization (M) step. Slightly different from regular EM, SSEM is based on stochastic sampling of latent variables in the E step in order to obtain (pseudo-)complete data. The word 'structural' refers to the fact that the M step involves a maximization not only over model parameters (parent weights and edge lengths) but also over the model structure (network topology). Inside the E step, an inner loop based on LBP replaces the variable elimination algorithm commonly used in dealing with latent variables in tree-structured phylogenies.

We initialize the structure as a phylogenetic tree obtained from by Neighbor-Joining algorithm [18], which is used for sampling of the latent variables in the first E step. After this the initial tree is discarded and in particular, it is not used to restrict the structure search in any way. The E and M steps are repeated until the objective function converges.

Let $o$ denote all the observed data, and let $L$ denote the latent variables. The model structure and parameters on each iteration, $t$, of the algorithm are denoted as $G^{(t)}$ and $\theta^{(t)} = (w^{(t)}, \tau^{(t)})$ respectively.

### 4.1 Stochastic E Step

Recall that in the E step, regular EM computes the expected complete data log-likelihood with respect to the latent variables

$$E_{L|o,G^{(t)},\theta^{(t)}} \left[ \log P_{(G,\theta)}(o, L) \right], \tag{3}$$

where the structure and parameters $G, \theta = (w, \tau)$ are allowed to differ from $G^{(t)}, \theta^{(t)}$. The above quantity is then maximized with respect to $G$ and $\theta$ in the M step. For complete data, under the i.i.d. assumption, the log-likelihood for a set of sequences of length $N$ becomes a sum with $N$ terms. We group them based on the configurations of a node and its parents:

$$\log P_{(G,\theta)}(o, l) = \sum_{\substack{i=1,\ldots,p \\ x \in \{A,C,G,T\} \\ pa \in \{A,C,G,T\}^{q_i}}} N_{ixpa} \log P_{\theta_i}(X_i = x \mid \mathrm{Pa}_i = pa), \tag{4}$$

where $q_i$ denotes the number of parents for variable $X_i$, and the count $N_{ixpa}$ indicates the number of sites where $X_i$ takes value $x$ and its parents take values $pa$. The counts $N_{ixpa}$ are called the *sufficient statistics* since given the model

parameters they uniquely determine the likelihood. For each combination of values $x, pa$, the conditional probability $P_{\theta_i}(X_i \mid \mathrm{Pa}_i)$ can be considered a constant, and the log-likelihood is a linear function of the sufficient statistics. Hence computing the *expected* log-likelihood only requires the computation of the expectation of the sufficient statistics, which can be done by summing over all the independent sites

$$E_{L|o,\theta^{(t)}}\left[N_{ixpa}\right] = \sum_{j=1}^{N} P_{(G^{(t)},\theta^{(t)})}(X_i^j = x, \mathrm{Pa}_i^j = pa \mid o^j),$$

where superscript $j$ indicates the site. The required computations may easily become infeasible since the conditional probabilities in the above formula may require complex inference procedures in case the network structure is not of a very specific kind (such as a tree).

Friedman et al. [6] suggest an approximation of the form

$$P_{(G^{(t)},\theta^{(t)})}(L \mid o) \approx \prod_{i=1}^{|L|} P_{(G^{(t)},\theta^{(t)})}(L_i \mid o), \tag{5}$$

where $|L|$ denotes the number of latent variables. This amounts to treating each node and its (potential) parent(s) as conditionally independent given the observed data. We suggest a different approximation which avoids the above drastic conditional independence assumption by sampling the latent variables from their conditional distribution $P_{(G^{(t)},\theta^{(t)})}(L \mid o)$. To do so, we exploit the chain rule

$$P_{(G^{(t)},\theta^{(t)})}(L \mid o) = \prod_{i=1}^{|L|} P_{(G^{(t)},\theta^{(t)})}(L_i \mid L_{1:i-1}, o) \tag{6}$$

We will sample a value $l_1$ for the latent variable $L_1$ from its LBP-approximated conditional distribution given the observed data $o$, after which we include the value $l_1$ in the set of (pseudo-)observed variables, and proceed recursively to sample all the remaining latent variables. The procedure is outlined as Algorithm 1 below.

**Data**: $o$: vector of observed data (at a single site)
**Result**: $l$: vector of sampled data for latent variables (at the same site)
**for** $i \in \{1, ..., |L|\}$ **do**
  Perform LBP to approximate $P_{(G^{(t)},\theta^{(t)})}(L_i \mid l_1, \ldots, l_{i-1}, o)$
  Draw value $l_i$ from the obtained distribution.
**end**
**return** $l_1, \ldots, l_{|L|}$

**Algorithm 1:** Sampling latent variables from their joint conditional distribution approximated by loopy belief propagation (LBP).

In practice, drawing a single sample vector, $l$, per site appears to be sufficient to obtain sufficiently accurate approximations of the expected counts unless the

number of sites is very small. This strategy is more generally called *stochastic* EM [2]. Theoretical and numerical results backing up its validity are presented in [14].

### 4.2 Structural M Step

Having sampled the latent variables in the E step to obtain the pseudo-complete data $(o, \tilde{l})$, the M step is used to estimate a phylogenetic hypothesis, i.e., the network structure, $G$, the weight parameters associated with possible reticulations, $w$, and the edge lengths, $\tau$. All of them are estimated by maximizing the following objective function:

$$(G^{(t+1)}, \theta^{(t+1)}) = \arg\max_{(G,\theta)} \log P_{(G,\theta)}(o, \tilde{l}), \tag{7}$$

where $\tilde{l}$ denotes the sampled values for all hidden variables obtained in the stochastic E step, and $\theta = (w, \tau)$. Any Bayesian network learning algorithm can be applied with the pseudo-complete data. We start with an empty network and apply local modifications including edge deletions, additions, and reversals until the likelihood score cannot be improved. Further heuristics including a tabu search to escape local optima are detailed in the next section.

The parameters can be estimated in a relatively straightforward manner under the JC model, which we use in our implementation, as well as other commonly used sequence evolution models.

### 4.3 Avoiding Local Optima and Overfitting

As is typical to EM-based algorithms, it is beneficial to implement some modifications that help to avoid the search from getting stuck to local optima. Since the method is based on maximizing the likelihood, it is also prone to overfitting unless some complexity regularization is performed.

First, to escape local optima in the structure search within the M step, we apply so called *tabu search* heuristic [7] where structure modifications that reduce the likelihood score are accepted in case there are no available local modifications that improve the score. To do so, we maintain a *tabu list* wherein we record recently visited graph structures in order to prevent repeatedly visiting the same structures. The search is terminated after a maximum number of iterations is reached or when no improvement in the best structure occurs in several steps, after which the overall best structure is returned.

Moreover, even if the M step finds the globally optimal structure given the pseudo-complete data $(o, \tilde{l})$, the EM iterations may end up in a local optimum of the incomplete-data likelihood, where the pseudo-observations sampled in the E step reinforce the current (locally optimal) structure hypothesis; see [6]. The stochastic EM algorithm is less prone to this problem than regular EM (see [14]) but when the sequence length is large enough, the problem persists. We therefore adopt the perturbation method in *deterministic annealing EM* by Ueda and

Nakano [24]. This means that the sampling distributions in Algorithm 1 is raised to power $\beta \leq 1$ and normalized after it has been inferred by LBP so that the pseudo-observations are drawn from a distribution proportional to

$$P_{(G^{(t)},\theta^{(t)})}(L_i \mid l_{1:i-1}, o)^{\beta}$$

where $1/\beta$ acts like a temperature parameter. The inverse temperature $\beta$ should be small at the beginning, so that the sampling distribution is close to uniform. When $\beta$ is increased, the distribution is perturbed less and it will approach the unperturbed distribution as $\beta \to 1$. Currently we heuristically set $\beta^{(1)} = 0.6$ and $\beta^{(t+1)} = \min\{1.0, 1.05\,\beta^{(t)}\}$.

Finally, to avoid overfitting due to the increased flexibility allowed by the reticulation nodes, we use the Bayesian information criterion (BIC) [19] to penalize the score function, which becomes

$$\mathrm{BIC}(G, \theta \mid o, \tilde{l}) = \log P_{(G,\theta)}(o, \tilde{l}) - \frac{k}{2} \log N, \tag{8}$$

where $k$ is number of free parameters in model $G$ (including both the weights and the edge lengths), and $N$ is the sequence length. The second term in the BIC score can be seen as a complexity penalty reducing the tendency to overfit. Because the penalizing term is the same in both complete and incomplete data, when BIC is used instead of ML as the scoring function in Eq. (7), the validity of the EM algorithm is maintained; see [3]. The good performance of BIC in preventing overfitting in phylogenetic networks has been observed by Park and Nakleh [15].

### 4.4 Postprocessing of the Networks

From the point of view of network search, the properties required from phylogenetic networks can be a problem since they might restrict the exploration of promising structures. Therefore, we perform the SSEM algorithm using unconstrained network structures, and apply the following sequence of postprocessing steps only after the algorithm has converged:

1. Recursively remove all unlabeled leaves.
2. Remove unlabeled nodes with in-degree and out-degree of 1.
3. Edge from two labeled nodes $(A, B)$ with length $\tau_{AB}$ is replaced by $(x, A)$ with $\tau_{xA} = \epsilon$ and $(x, B)$ with $\tau_{xB} = \tau_{AB}$, where $x$ is an internal node and $\epsilon \approx 0$.
4. An internal node $x$ with more than two children, $x_1, x_2, ...$, is replaced by a new internal node $y$ with children $x_1$ and $x$, and $x_1$ is removed from the children of $x$. This rule is applied recursively until $x$ has at most two children.

We refer the reader to [6] for detailed illustrations and the proof why these alterations do not change the score.

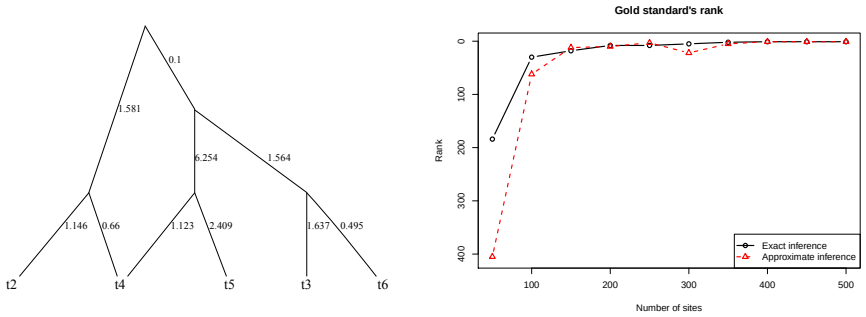


**Fig. 1.** *Left:* The true phylogenetic network. Edge lengths (shown along the edges) are drawn from an exponential distribution. *Right:* Ranks of the true network as a function of sequences length using exact and approximate computations. Both curves tend to increase which means that as the sample size grows, they eventually rank the true network first.

## 5 Experiments

To demonstrate the practical utility of the proposed method, we perform experiments on both simulated and real data. We first illustrate the accuracy of the likelihood evaluation based on loopy belief propagation which we use in the E step to show that model comparison based on approximated likelihood is reliable. We then demonstrate the PhyloDAG method by applying it to both simulated and real data.

### 5.1 Exact vs Approximated Likelihood: an Illustration

We apply a procedure where we simulate DNA data with increasing sequence length, $N = 50, 100, 150, \ldots, 500$ for the leaf nodes of an arbitrary *tree* structure following the JC model and no insertions or deletions (indels). To create a hybrid node, we pick two leaf nodes and produce a hybrid sequence by randomly copying the character at each site from either one of chosen the leaf nodes according to some fixed weights. The two leaf nodes are then removed and replaced by the hybrid node whose parents are those of the removed leaf nodes. The resulting phylogenetic network is shown in Fig. 1.

We then modify the true structure by adding and removing edges to obtain a sample of 1000 incorrect topologies (including some duplicates). We rank these 1001 phylogenies by their BIC scores where instead of the pseudo-complete likelihood $P(o, \tilde{l})$ we use the incomplete-data likelihood $P(o)$ so that the scores are comparable across different networks. We compare the ranking performance obtained by using an exact brute-force computation of the incomplete-data likelihood and the LBP approximation. Since the problem size is small, even the exact computation takes less than three seconds for samples up to $N = 500$ using an efficient implementation. In the case of LBP, we use the identity $P(o) =$

$P(\tilde{l}, o)/P(\tilde{l} \mid o)$ which holds for all $\tilde{l}$. The LBP approximation takes less than 0.5 seconds. Since the exact computation takes exponential time in the number of latent variables, it quickly becomes useless in practice as the problem size is increased, whereas the LBP method scales to much bigger problems.

In Fig. 1, the ranks of the true phylogenetic network by both exact and approximate inference are plotted against the sample size. In both methods the rank of the true structure tends to improve as the sequence length grows. The brute-force method ranks the true structure higher for sequences up to 100 nucleotides but for longer sequences the differences are generally very small.

### 5.2 Structure Search on Synthetic Data

Following the experimental procedure described above, we generate a data set with 15 taxa and sequence length 2000. The true underlying phylogenetic network is shown in Fig. 2. We apply PhyloDAG as well as PhyloNet[2], a recent method proposed by Yu et al. [27].

Figure 3 shows the result of PhyloDAG. In order to make it easier to compare the structure inferred by PhyloDAG to the correct network, four groups of taxa are shaded and labelled as $A$–$D$. Except some minor differences like the position of group $C$ (taxa $t6$ and $t17$), PhyloDAG infers the structure almost correctly. In particular, the two reticulation events at $t7$ and $t9$ are inferred correctly. Note that the BIC criterion was used to decide the number of reticulate edges in the model based on the data without user intervention.

In PhyloNet, we apply the maximum likelihood phylogenetic network method. PhyloNet requires a backbone tree, and as suggested by Yu et al. [27], we use a backbone obtained by MrBayes [17]. PhyloNet also requires that the number of reticulations be specified, and we provide the correct number, two. Other settings of PhyloNet are set to default values. By default, the algorithm is repeated 10 times and the network that maximizes the likelihood as computed by PhyloNet is produced as the output.

Figure 4 shows the structure inferred by PhyloNet. The solid edges are from the backbone tree by MrBayes and dotted edges are the added reticulation edges. In this experiment, despite the good backbone tree, the two reticulation edges suggested by PhyloNet are incorrect. The reticulation edge near $t11$ may correspond to an actual reticulation (see Fig. 2) between the immediate ancestors of $t11$ and $t4$ which results in the sequence at $t7$ but it is still relatively far from correct. It will be interesting to analyse in detail why PhyloNet produces reticulate edges between neighboring nodes only. The experiments presented by Yu et al. [27] do not test whether this behavior occurs generally: they involve only 4 or 5 taxa so that reticulation between more distant branches cannot be investigated. Another possible explanation for the poor result is a different sequence evolution model employed in PhyloNet whereas PhyloDAG may benefit from the fact that it is based on the JC model which is also used to simulate the sequences – however, see the results on real data in the next subsection.

[2] `http://bioinfo.cs.rice.edu/phylonet`

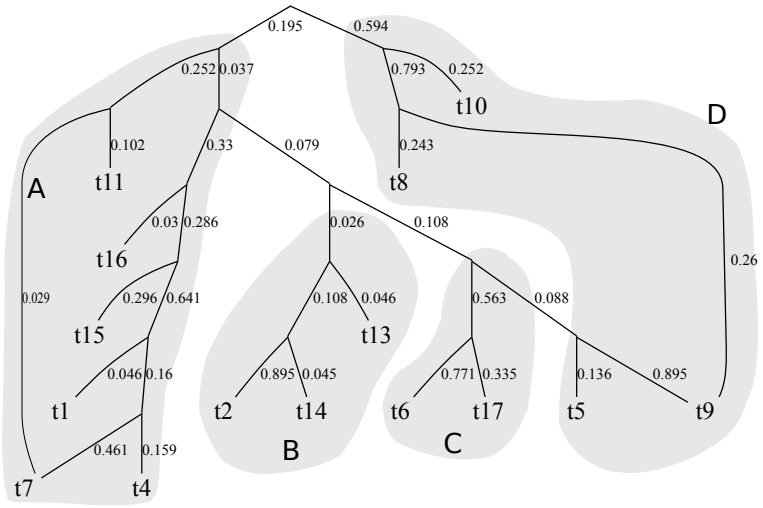


**Fig. 2.** The true phylogenetic network used to simulate 15 sequences, including two reticulations (taxa $t7$ and $t9$). Numbers indicate edge lengths. The groups $A$–$D$ are shaded for clarity.

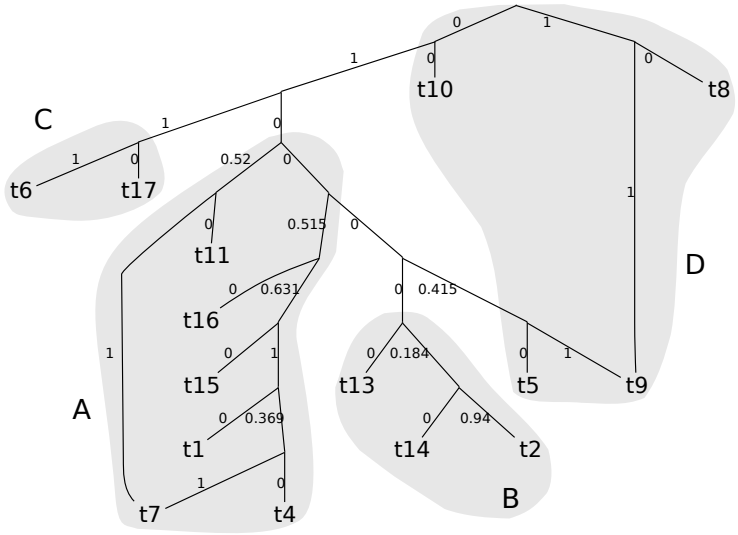


**Fig. 3.** Result of PhyloDAG for data simulated from the network in Fig. 2. Numbers indicate estimated edge lengths.

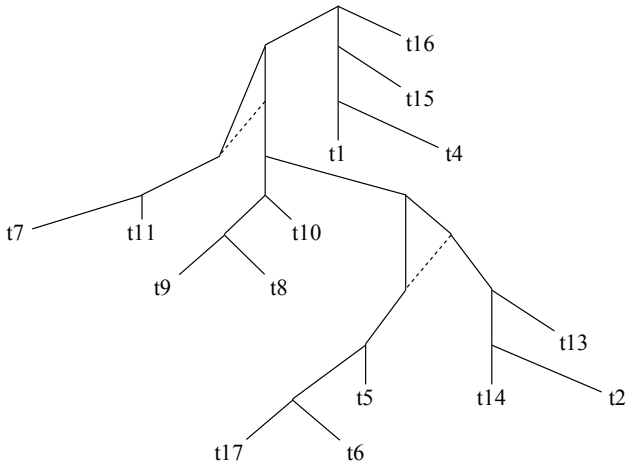


**Fig. 4.** Result of PhyloNet [27] for data simulated from the network in Fig. 2.

The test is done on an 3.4 GHz 8 core CPU computer with 16 GB of memory. For this data set, PhyloDAG runs 12 iterations of the SSEM procedure which takes less than three minutes. On the same setup, PhyloNet runs for about five hours (excluding the running time of MrBayes).

### 5.3 Real Data Experiment

We test PhyloDAG on a real data set "Feliner"[3]. This is one of a few data sets where the underlying phylogenetic network is at least partially known since they result from an artificial hybridization of *Armeria* plants in a greenhouse [1].

The data contains a number of *Armeria villosa ssp. longiaristata* (VIL) and *Armeria colorata* (COL) plants. The specimens VIL#58/120 and COL#11/12 were crossed to create a hybrid generation labeled F1. We select a subset of the original data set that includes hybrid taxa and their ancestors, so that the relationships between the taxa are known from the experiment and the results are easy to interpret. We expand heterozygous sites as pairs of nucleotides following the encoding of Aguilar et al. [1] (for example, $W$ in the sequence is expanded as nucleotides $AT$). The total sequence length is 626 nucleotides after the preprocessing. The problem is complicated by the fact that all the sequences are very similar to each other: they differ at not more than 10 sites.

Figure 5 shows the results of PhyloDAG on the subset of seven sequences from the Feliner data. PhyloDAG groups the COL and VIL families correctly and includes a reticulation edge correctly identifying the hybrid ancestry of the F1 family. The edge lengths are compatible with the observation that the

[3] `http://www.rjr-productions.org/Database.html`

F1 sequences are very close to the COL sequences (about 4–5 differences) and somewhat less close to the VIL sequences (about 7–10 differences).

Figure 6 shows the PhyloNet result, obtained using default settings. The backbone tree (solid lines) obtained by MrBayes places the hybrid F1 species between the ancestor groups COL and VIL. The PhyloNet method was repeated twice: first, setting the number of reticulations to one, and another time, setting it to two. The network show in the figure includes all the reticulate edges (dotted lines) appearing in either of the the resulting networks. Similar to the simulation experiment, the reticulate edges by PhyloNet are near the hybrid taxa but their end points are too close to each other to provide useful information about the ancestry of the hybrids.

## 6 Conclusions

We propose a new method, PhyloDAG, for constructing likelihood based phylogenetic networks from sequence data. The method is based on *i*) structural EM which treats the graph structure as a parameter to be optimized in the M step *ii*) an efficient stochastic implementation of the E step based on loopy belief propagation. The key difference in the procedure compared to earlier likelihood-based approaches is that whereas earlier methods tend to involve an EM or Monte Carlo type algorithm as an inner loop of a structure learning process, we put the structure learning procedure inside the M step of an EM-type algorithm. This significantly speeds up the structure learning process since it avoids costly iterative likelihood evaluations, and allows an unrestricted structure search without a fixed backbone tree.

We presented simulations and a real data experiment to demonstrate the accuracy of the method. Compared to another recent likelihood-based method, PhyloDAG was orders of magnitude faster and produced much more accurate network structures. Variations of our method can be constructed where different models of reticulation are applied. Additional large scale experiments with real and simulated data will be required to assess the benefits of our approach.

## Acknowledgments

This work was supported in part by the Academy of Finland (Center-of-Excellence COIN). We are grateful to Vincent Moulton for insightful comments. The anonymous reviewers suggested a comparison to the PhyloNet method and made several other suggestions that significantly improved the paper.

## References

1. Aguilar, J.F., Rosselló, J., Feliner, G.N.: Nuclear ribosomal DNA (nrDNA) concerted evolution in natural and artificial hybrids of Armeria (Plumbaginaceae). Molecular Ecology 8(8), 1341–1346 (1999)

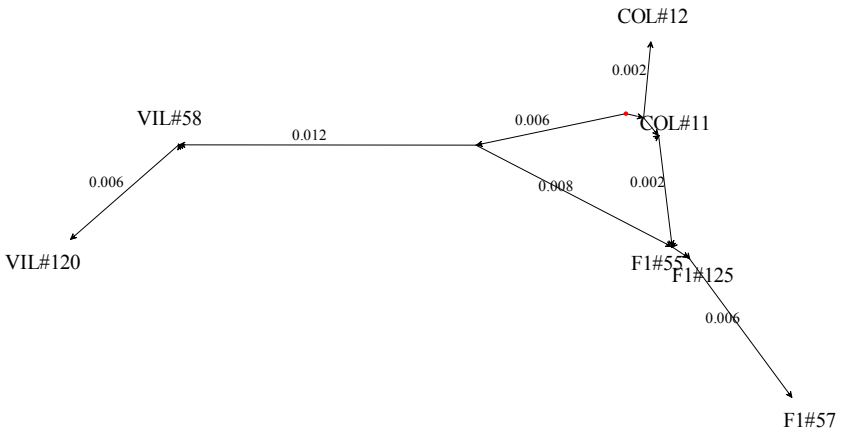


**Fig. 5.** PhyloDAG result for the Feliner data. In this case, edge lengths are drawn proportional to their estimates.

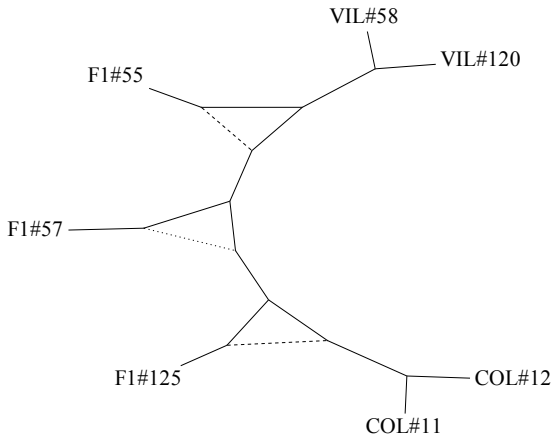


**Fig. 6.** PhyloNet result for the Feliner data. (Edge lengths not drawn proportional to their estimates.)

2. Celeux, G., Diebolt, J.: The SEM algorithm: A probabilistic teacher algorithm derived from the EM algorithm for the mixture problem. Computational statistics quarterly 2(1), 73–82 (1985)
3. Dempster, A.P., Laird, N.M., Rubin, D.B.: Maximum likelihood from incomplete data via the EM algorithm. Journal of the Royal Statistical Society. Series B 39(1), 1–38 (1977)
4. Durbin, R., Eddy, S.R., Krogh, A., Mitchison, G.: Biological Sequence Analysis: Probabilistic Models of Proteins and Nucleic Acids. Cambridge University Press (1998)
5. Felsenstein, J.: Evolutionary trees from DNA sequences: A maximum likelihood approach. Journal of Molecular Evolution 17, 368–376 (1981)
6. Friedman, N., Ninio, M., Pe'er, I., Pupko, T.: A structural EM algorithm for phylogenetic inference. Journal of Computational Biology 9(2), 331–353 (2002)
7. Glover, F., Laguna, M.: Tabu Search. Kluwer Academic Publishers, Norwell, MA (1997)
8. Haeseler, A., Churchill, G.A.: Network models for sequence evolution. Journal of Molecular Evolution 37(1), 77–85 (1993)
9. Husmeier, D., Wright, F.: Detection of recombination in DNA multiple alignments with hidden Markov models. Journal of Computational Biology 8(4), 401–427 (2001)
10. Jin, G., Nakhleh, G., Snir, S., Tuller, T.: Maximum likelihood of phylogenetic networks. Bioinformatics 22, 2604–2611 (2006)
11. Jukes, T.H., Cantor, C.R.: Evolution of protein molecules. Mammalian protein metabolism 3, 21–132 (1969)
12. Meng, C., Kubatko, L.S.: Detecting hybrid speciation in the presence of incomplete lineage sorting using gene tree incongruence: a model. Theoretical population biology 75(1), 35–45 (2009)
13. Morrison, D.: Introduction to Phylogenetic Networks. RJR Productions (2011)
14. Nielsen, S.F.: The stochastic EM algorithm: estimation and asymptotic results. Bernoulli pp. 457–489 (2000)
15. Park, H.J., Nakhleh, L.: Inference of reticulate evolutionary histories by maximum likelihood: the performance of information criteria. BMC Bioinformatics 13(Suppl 19), S12 (2012)
16. Pearl, J.: Probabilistic Reasoning in Intelligent Systems: Networks of Plausible Inference. Morgan Kaufmann Publishers Inc., San Francisco, CA, USA (1988)
17. Ronquist, F., Huelsenbeck, J.P.: MrBayes 3: Bayesian phylogenetic inference under mixed models. Bioinformatics 19(12), 1572–1574 (2003)
18. Saitou, N., Nei, M.: The neighbor-joining method: a new method for reconstructing phylogenetic trees. Molecular Biology and Evolution 4, 406–425 (1987)
19. Schwarz, G.: Estimating the dimension of a model. The Annals of Statistics 6(2), 461–464 (1978)
20. Sneath, P.H.A.: Cladistic representation of reticulate evolution. Systematic Zoology 24, 360–368 (1975)
21. Strimmer, K., Moulton, V.: Likelihood analysis of phylogenetic networks using directed graphical models. Molecular Biology and Evolution 17(6), 875–881 (2000)
22. Strimmer, K., Wiuf, C., Moulton, V.: Recombination analysis using directed graphical models. Molecular Biology and Evolution 18(1), 97–99 (2001)
23. Tehrani, J., Nguyen, Q., Roos, T.: Oral fairy tale or literary fake? Investigating the origins of Little Red Riding Hood using phylogenetic network analysis. Digital Scholarship in the Humanities (2015), to appear
24. Ueda, N., Nakano, R.: Deterministic annealing EM algorithm. Neural Networks 11(2), 271–282 (1998)

25. Webb, A., Hancock, J.M., Holmes, C.C.: Phylogenetic inference under recombination using Bayesian stochastic topology selection. Bioinformatics 25(2), 197–203 (2009)
26. Whelan, S., Lio, P., Goldman, N.: Molecular phylogenetics: state-of-the-art methods for looking into the past. TRENDS in Genetics 17(5), 262–272 (2001)
27. Yu, Y., Dong, J., Liu, K.J., Nakhleh, L.: Maximum likelihood inference of reticulate evolutionary histories. Proceedings of the National Academy of Sciences 111(46), 16448–16453 (2014)