

Likelihood Analysis of Phylogenetic Networks Using Directed Graphical Models

Korbinian Strimmer*¹ and Vincent Moulton†

*GSF-Forschungszentrum für Umwelt und Gesundheit, MIPS, am Max-Planck-Institut für Biochemie, Martinsried, Germany; and †FMI, Physics and Mathematics Department, Mid Sweden University, Sundsvall, Sweden

A method for computing the likelihood of a set of sequences assuming a phylogenetic network as an evolutionary hypothesis is presented. The approach applies directed graphical models to sequence evolution on networks and is a natural generalization of earlier work by Felsenstein on evolutionary trees, including it as a special case. The likelihood computation involves several steps. First, the phylogenetic network is rooted to form a directed acyclic graph (DAG). Then, applying standard models for nucleotide/amino acid substitution, the DAG is converted into a Bayesian network from which the joint probability distribution involving all nodes of the network can be directly read. The joint probability is explicitly dependent on branch lengths and on recombination parameters (prior probability of a parent sequence). The likelihood of the data assuming no knowledge of hidden nodes is obtained by marginalization, i.e., by summing over all combinations of unknown states. As the number of terms increases exponentially with the number of hidden nodes, a Markov chain Monte Carlo procedure (Gibbs sampling) is used to accurately approximate the likelihood by summing over the most important states only. Investigating a human T-cell lymphotropic virus (HTLV) data set and optimizing both branch lengths and recombination parameters, we find that the likelihood of a corresponding phylogenetic network outperforms a set of competing evolutionary trees. In general, except for the case of a tree, the likelihood of a network will be dependent on the choice of the root, even if a reversible model of substitution is applied. Thus, the method also provides a way in which to root a phylogenetic network by choosing a node that produces a most likely network.

Introduction

Probabilistic data modeling is one of the most powerful and efficient approaches to molecular sequence analysis (Durbin et al. 1998). Among the many virtues of statistical model building is the possibility of assigning likelihoods that allow one to evaluate and discriminate competing hypotheses for how an observed data set arose. In phylogenetics, likelihood methods were applied first to gene frequency data (Edwards and Cavalli-Sforza 1964) and subsequently also to molecular sequences (Neyman 1971; Kashab and Subas 1974). However, it was not until Felsenstein's (1981) work that a general and practical procedure for calculating the likelihood of a set of sequences related by a hypothesized evolutionary tree became available.

In recent years, it has become evident that many data sets from fields as diverse as epidemiology (Holmes, Worobey, and Rambaut 1999), population genetics (Oota et al. 1999), genomics (Lake, Jain, and Rivera 1999), early evolution (Doolittle 1999), etc., exist for which the evolution of the sequences cannot properly be represented by a tree. Due to recombination and horizontal gene transfer, multiple sources may contribute to a single gene. In these instances, the underlying structure of the data can be represented by phylogenetic networks (Bandelt 1994; Dress, Huson, and Moulton 1996). These networks extend the concept of trees and can be

viewed in part as a combination of different treelike histories. Algorithms for constructing phylogenetic networks have been developed on the basis of both pairwise distances and parsimony criteria (Bandelt and Dress 1992, 1993; Bandelt et al. 1995). von Haeseler and Churchill (1993) provided a framework for evaluating likelihoods of binary sequences related by networks. However, a general procedure for statistically assessing a phylogenetic network is still lacking.

Here we present a likelihood approach for arbitrary phylogenetic networks based on directed graphical models. Our approach generalizes the approach of Felsenstein (1981), including it as a special case when the network in question is a tree. This enables the direct comparison of both trees and networks on a statistically sound basis. Furthermore, given suitable network search procedures, it opens the way to inferring maximum-likelihood networks.

The rest of the paper is organized as follows. The next section gives an overview of directed graphical models, a powerful framework for describing complex systems of statistically correlated random variables. Next, we show how the approach of Felsenstein (1981) can be conveniently restated in terms of graphical models of sequence evolution and how the likelihood of a tree can be obtained within this framework using a variant of Markov chain Monte Carlo sampling. Subsequently, the model is generalized to arbitrary phylogenetic networks. Finally, using this method, we investigate a viral data set.

Directed Graphical Models

Graphical models provide a marriage between probability theory and graph theory (Lauritzen 1996). Essentially, they are graphs in which nodes represent stochastic variables, and edges between nodes indicate

¹ Present address: Department of Zoology, University of Oxford, Oxford, England.

Key words: maximum likelihood, phylogenetic network, graphical model, Bayesian network, evolutionary tree, Markov chain Monte Carlo sampling.

Address for correspondence and reprints: Vincent Moulton, FMI, Physics and Mathematics Department, Mid Sweden University, S 851-70, Sundsvall, Sweden. E-mail: vince@dirac.nts.mh.se.

Mol. Biol. Evol. 17(6):875–881. 2000

© 2000 by the Society for Molecular Biology and Evolution. ISSN: 0737-4038

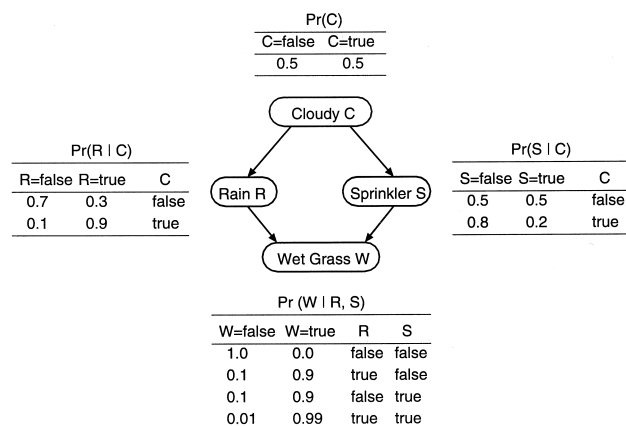


FIG. 1.—Bayesian network describing probabilistic dependences among variables Cloudy, Sprinkler, Rain, Wet Grass.

correlation between these variables. If the graph is directed, then the edges represent conditional dependencies, and one has what is known as a directed graphical model. These models are also known as Bayesian networks, influence diagrams, or belief networks (Russell and Norvig 1995). Graphical models constitute a modular high-level language that allow one to explicitly represent the dependences or independences between variables while ignoring numerical details. At the same time they provide a clear framework for quickly deriving any underlying probabilistic equation of interest and for computing corresponding probability distributions. The theory of graphical models is highly developed (Jordan 1999; Pearl 1988). They have been applied to numerous problems in applied mathematics and engineering, especially in machine learning and artificial intelligence, where graphical models encode uncertain knowledge in expert systems. An annotated bibliography on these models has been collected by Buntine (1996), whereas Krause (1998) provides an accessible tutorial.

We now consider a simple (and often used) example of a Bayesian network (e.g., figure 2.1.2 in Jensen 1996) to illustrate some of the relevant features of these objects. In figure 1, four nodes representing the variables Cloudy (*C*), Sprinkler (*S*), Rain (*R*), and Wet Grass (*W*) are connected to form a so-called directed acyclic graph (DAG). Note that directed cycles are disallowed in a Bayesian network so as to prohibit the possibility of a node influencing itself through a chain of intermediate nodes. The variables *C*, *R*, *S*, and *W* take on one of two states (true, false). Also, each node is assigned a probability of observing a state at the node given the states of the parent nodes, which we give in a table next to the node. These probabilities form the basis for computing the joint probability

$$\begin{aligned} \Pr(C, R, S, W) \\ = \Pr(C)\Pr(R|C)\Pr(S|C)\Pr(W|R, S), \end{aligned} \quad (1)$$

i.e., the probability of observing a combination of states assuming the specified probabilistic dependencies among the nodes. Thus, in this example, the probability

of seeing wet grass for a cloudy sky with rain and no sprinkler on is $0.5 \times 0.9 \times 0.8 \times 0.9 = 0.324$.

Not all states in the network may be known or observable at any given time. In this case, marginal probabilities for subsets of known states are obtained from the joint probability by summing over all possible combinations of states for the unknown variables. For example, if we could not observe *R* or *S*, we would obtain the marginal probability

$$\Pr(C, W) = \sum_R \sum_S \Pr(C, R, S, W), \quad (2)$$

i.e., the likelihood of the observed states, taking into account dependences among hidden variables. In this way, the unknown variables *R* and *S* are eliminated from the distribution. Thus, in our example, we see that the likelihood of seeing wet grass when it is cloudy is $(0.5 \times 0.1 \times 0.8 \times 0.0) + (0.5 \times 0.1 \times 0.2 \times 0.9) + (0.5 \times 0.9 \times 0.8 \times 0.9) + (0.5 \times 0.9 \times 0.2 \times 0.99) = 0.4221$.

Computing the marginal probabilities generally takes exponential time in the number of unknown nodes, as the number of terms in equation (2) increases exponentially in the unknown variables. However, for special kinds of networks such as trees, there exist economies for computing these probabilities based on variable elimination (Horner's rule) that rearrange the individual terms in equation (2) for more efficient computation. Unfortunately, in the general case, no simple savings of this kind apply. Instead, it is usually necessary to apply approximations, for example, Monte Carlo simulations (Chickering and Heckerman 1997; Jordan 1999). Alternatively, it is also possible to convert the belief network into a secondary treelike structure (junction tree) and then compute probabilities working with this structure (Jensen 1996). However, the difficulty of finding an optimal junction tree remains.

Directed Graphical Models and Sequence Evolution on Trees

Felsenstein (1981) introduced a framework for the calculation of the likelihood of a set of DNA sequences given a tree as an evolutionary hypothesis. As we now see, it is straightforward to translate his framework into a directed graphical model.

The underlying DAG is provided by rooting the tree at a prescribed node and directing all edges in the tree away from this node (see fig. 2a). The nodes of this DAG are considered as variables with four nucleotide states, and directed edges introduce dependences among these variables. In particular, each node of the network is assigned a nucleotide state *x*, *y*, *z*, . . . , where states of internal nodes (which usually correspond to unknown sequences) are ancestral to the states observed at the external nodes (which correspond to observed sequences).

The directed graphical model also requires the prescription of local node probabilities that specify the probability of observing state *x* at a selected node given

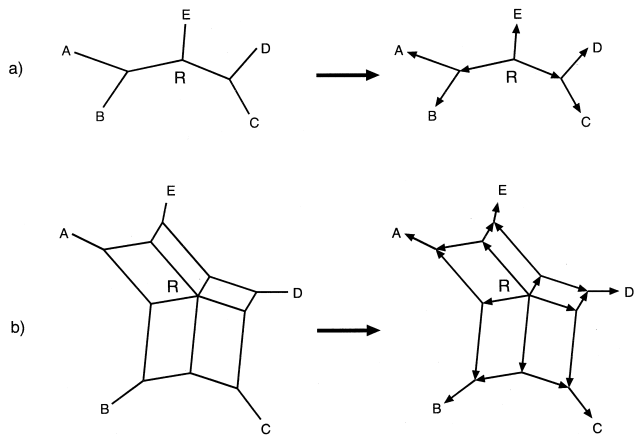


FIG. 2.—Rooting (a) a tree and (b) an outer-planar graph at node R to obtain a directed acyclic graph. Note that parallel edges all point in the same direction, away from R.

the states of its direct parents. There are a variety of models of sequence evolution that are suited to this task (Liò and Goldman 1998). Describing the nucleotide substitution along an edge by a Markov process, they assign a transition probability $P_{xy}(k)$ that gives the probability of starting in state x and observing state y after k substitutions. Note that in these models, $P_{xy}(k)$ incorporates corrections for multiple hits, and when k is large, the stationary distribution is reached, i.e., $P_{xy}(k) = \pi_y$ as $k \rightarrow \infty$. Now, in the case in which a node has no direct parent, i.e., it is the root node, the probability $\Pr(x)$ of observing x equals π_x (fig. 3a). If a node has a single direct parent (fig. 3b), the probability $\Pr(x|y)$ of observing x given the parent state y is the transition probability $P_{yx}(k_y)$, where k_y denotes the length of the branch to the parent state y .

The likelihood of the observed sequences assuming the tree is then calculated using the obvious generalization of equation (2). Note that this involves only a single column of a sequence alignment. To obtain the likelihood of a complete alignment, the likelihoods of all sites are multiplied together, assuming independence of sites (Felsenstein 1981; Felsenstein and Churchill 1996). To reduce the complexity of the summation involved in computing the likelihood, Felsenstein applied a variable elimination scheme (“pruning”) based on using a post-order traversal of the nodes in the tree. For an arbitrary DAG, however, there is no obvious generalization of this procedure. In the present paper, we resort to Monte Carlo simulations to compute the likelihood. This class of approximations is known to provide accurate results (Chickering and Heckerman 1997). Unfortunately, the amount of computer time needed for convergence can still be very large. We are currently exploring alternative and more efficient computation schemes applying other approximation techniques and using, e.g., junction trees that will be published elsewhere. Here we will briefly review Gibbs sampling (Gelfand and Smith 1990), a simple stochastic procedure that we will employ in this paper.

Luckily, it turns out that only a minor fraction of the terms in the likelihood summation (eq. 2) contribute

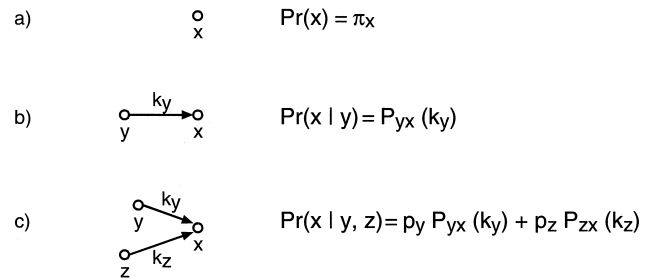


FIG. 3.—Components of the directed graphical model used to compute the likelihood of data related by a phylogenetic network. Formulas are explained in the text.

significantly to the overall likelihood, thus making it possible to approximate the likelihood by summing only these terms. To detect the combination of states at the unknown nodes that contribute most to the likelihood, a search procedure that visits the most important combinations is necessary. Stochastic procedures like Gibbs sampling are suited for this purpose. Basically, we start from a random assignment of nucleotide states at the unknown nodes, computing its probability using the joint probability distribution (eq. 1). Subsequently, one of the internal nodes is selected at random, and its current state is replaced by drawing a new state from its distribution conditional on the current states of all the other nodes, again using equation (1). This replacement procedure is repeated as often as desired. It is possible to show (Gelfand and Smith 1990) that in this way a series of combinations of assignments of states to the internal nodes is generated, such that their frequencies in the chain are proportional to their importances in the sum in equation (2). Thus, all important combinations of states are eventually visited. They are stored for computation of the overall likelihood.

Before going on, we should make it clear that although the above discussion on directed graphical models was based on DNA sequences, it can also be applied to protein sequences, making the appropriate changes where necessary.

Likelihoods for Phylogenetic Networks

Considering the previous discussion on trees, it is clear that we could, in theory at least, assign likelihoods to a data set related by any given network representing an evolutionary hypothesis, as long as we could find a way to convert it into a directed graphical model. However, rather than considering this process in full generality, we concentrate on computing likelihoods for phylogenetic networks, a special class of networks that have been successfully used in phylogenetic analysis which are amenable to this technique.

Given a set of taxa, it is natural to consider bipartitions or splits of the taxa. For example, if the taxa A, B, C, D, E are analyzed, it may turn out that there is clear evidence (coming from, for example, distance or parsimony considerations) for separating the groups A, B, E and C, D . Phylogenetic networks are graphs that can be used to represent collections of splits derived from the sequences in question (Bandelt 1994). If a set

Table 1
Classes of Phylogenetic Networks

Network	Properties	Constraint on Splits	Maximum No. of Splits
Tree	Planar, each node has exactly one direct parent	Must be compatible	$2n - 3$
Outer-planar graph	Planar, nodes have one or more direct parents	Must be circular	$\binom{n}{2}$
Median graph	In general nonplanar, nodes have one or more direct parents	No constraint	$2^{n-1} - 1$

NOTE.—The maximum number of splits includes the n trivial splits. See also figure 4 and appendix B.

of splits derived from the taxa is compatible, the relationship among the taxa can be represented by a tree (Buneman 1971), and so trees are in particular examples of phylogenetic networks. However, if this is not the case, then phylogenetic networks represent sets of contradicting splits by hypercubes. Note that this is in contrast to consensus procedures in which contradictory or incompatible splits are dropped in order to arrive at a tree structure (Margush and McMorris 1981). As a result, phylogenetic networks can be rather complex objects falling somewhere between the extremes of being a tree or a hypercube.

In table 1 and figure 4, we present some simple phylogenetic networks. Note that a collection of parallel edges in a given network corresponds to a split represented by the network. Moreover, parallel edges are all assumed to be assigned the same length. In case the set of splits derived from the data is circular—as is quite often the case for molecular data when using the program SplitsTree (Huson 1998)—the relationship between the taxa can be visualized by a planar phylogenetic network, called an outer-planar graph (A. W. M. Dress and D. H. Huson, personal communication). Please refer to appendix B for a brief review of these terms.

We now show how a directed graphical model can be constructed for an outer-planar graph in a way that naturally generalizes the one that we used for trees, noting also that this method can be performed for arbitrary phylogenetic networks.

We begin by turning our given outer-planar graph into a DAG. To do this, any vertex is selected as a root node, and all edges are directed away from this node (see fig. 2*b*). It can be shown that in this way we produce a DAG in which (1) the root node is the only node that is a source, (2) the terminal nodes are the only nodes that are sinks, and (3) parallel edges are assigned the same direction. Note that from the biological point of view, it not only makes sense to have a unique root

node, but assigning the same direction to parallel edges having—per definition—the same length is also plausible. In particular, the direction of an edge indicates the flow of information from one sequence to another, and for a given split, information flows uniformly from one group of related sequences to the other.

We now explain how to assign local node probabilities to the DAG we have obtained. In a network, a node can have several direct parents. Therefore, we need to extend the assignment of local node probabilities given for a tree to the case of two or more parents (fig. 3*c*). We present the case with two parents, the generalization to three or more being clear. We define the probability $\Pr(x|y, z)$ of observing state x given states y and z to be $p_y P_{yx}(k_y) + p_z P_{zx}(k_z)$, where p_y, p_z denote the prior probabilities that the node in state x is influenced by the direct parent in state y, z , respectively. This choice of assignment of probabilities has the advantage of generalizing easily to an arbitrary number of parents and of having a clear Bayesian interpretation.

The prior probabilities are also easily interpreted as recombination parameters. For a sequence of a given length, they denote the proportion of sites stemming from one of the parents. We note that these parameters are not derived from the underlying phylogenetic graph. In contrast, they are to be estimated from the sequence data. As a consequence, the maximum-likelihood value for a given data set related by a network is always at least as high as the largest likelihood value of a tree embedded in the network. This can be seen by appropriately setting the prior probabilities to 0 or 1. If estimation of priors is not feasible and no additional information on the preference of parents is available, we suggest using uniform priors ($p_y = p_z = 1/2$). This is reasonable not only because the strength of the correlation is already described by the branch lengths k_y and k_z , but also because the two parents share a common ancestor themselves, and the distances of the node in state x to that common ancestor through any path involving one of the parents are identical, as parallel edges in the network are of equal length. Thus, no path is preferred to arrive at the common ancestor.

The likelihood of a network can then be computed by marginalization of the joint probability (eq. 2), using, e.g., MCMC sampling to approximate the sum. Note that in contrast to the special case of an evolutionary tree (Felsenstein 1981), even if a reversible model of substitution is used so that $\pi_x P_{xy}(k) = \pi_y P_{yx}(k)$ applies,

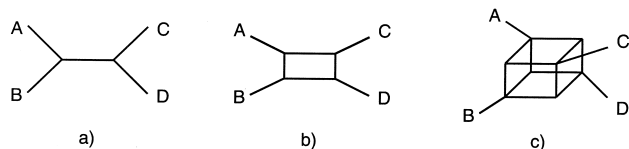


FIG. 4.—Examples of phylogenetic networks for taxa A, B, C, D. (a) Tree. (b) Outer-planar graph. (c) Median network. Note that parallel edges, i.e., edges corresponding to the same split, have equal lengths.

Table 2
The Four Best Trees for the HTLV Data Set

Tree	Log Likelihood
((L76045, L26585), L76050, (L76054, D13784))	-1,505.88
((L76045, L76050), L26585, (L76054, D13784))	-1,507.65
((L76045, L76050), L76054, (L26585, D13784))	-1,508.66
((L76050, L76054), L76045, (L26585, D13784))	-1,516.64

the likelihood can vary depending on the location of the root (a fact that follows from the construction of the local node probabilities and the sum used for computing the likelihood).

Before we close this section, we end with a word of caution. Even though it is stated above that we could possibly use this approach for computing the likelihood of arbitrary networks, this could, in general, lead to several problems. For example, even if we drop the constraint that parallel edges should have equal lengths for a phylogenetic network, problems with multiple maxima of the likelihood function can occur (data not shown). Moreover, in the general case, it is not at all clear how networks should be unambiguously rooted or how prior probabilities should be assigned to parent nodes.

HTLV Data

Using the method presented in the previous sections, we analyzed a data set consisting of five nucleotide human T-cell lymphotropic virus (HTLV) sequences of 930 sites. The data encode a fragment of the envelope surface glycoprotein 46 (GP46) gene (GenBank accession numbers L76054, L76050, D13784, L76045, and L26585). Using PUZZLE, version 4.0.2 (Strimmer and von Haeseler 1996), estimates for the parameters of the Tamura-Nei model (Tamura and Nei 1993) were obtained (transition/transversion parameter $\kappa = 12.52$, Y/R transition parameter $\tau = 0.35$), and a maximum-likelihood distance matrix was computed. Subsequently, the 15 possible tree topologies for five sequences were evaluated with PUZZLE. Using the distance matrix, a phylogenetic network was inferred using the program SplitsTree, version 2.1.1 (Huson 1998).

The list of the four best trees, along with their corresponding likelihood values, is shown in table 2. The outer-planar network is displayed in figure 5. Note that each of the four best trees is contained (as a subnetwork) in the network. As a basis for computing the likelihood for a network of the sequences and for inference of corresponding maximum-likelihood branch lengths and recombination parameters, the network shown in figure 6 was selected. This simplified network captures the essential nontreelikeness of the data while focusing on the strongest phylogenetic signal in the data. Its simplicity also helped us to perform the necessary computation and optimization in short time on a microcomputer.

To assess the accuracy of the Monte Carlo simulations we first recomputed the likelihood of the maximum-likelihood tree ($\log L = -1,505.88$) using Gibbs sampling (chain length 1,200) and recovered exactly the same log likelihood calculated by PUZZLE. Then we

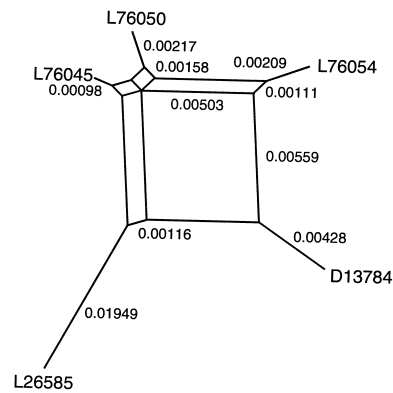


FIG. 5.—Split decomposition network for the HTLV data set. Sequences are labeled by their GenBank accession numbers. Branch lengths are given in nucleotide substitutions per site.

used the same technique to compute the maximum likelihood for the network using sequence L7650 as out-group and placing the root at node R1 (fig. 6). After optimizing the branch lengths and the prior probability, a log likelihood of $-1,500.32$ was obtained. Therefore, a difference in likelihood of several orders of magnitude was observed, favoring the maximum-likelihood network over the competing tree ($\log L = -1,505.88$). This is astonishing, as the network graph selected for the computation does not include the maximum-likelihood tree but only the second- and third-best trees. Unfortunately, a comparison using a Kishino-Hasegawa-Templeton test (Templeton 1983; Kishino and Hasegawa 1989) could not detect a significant difference in likelihood on a 5% basis between the network and the maximum-likelihood tree. However, this is probably not surprising, given that the four best trees are also not significantly different and that the data set contains only 49 variable positions (4.2% of all sites).

As the likelihood of a phylogenetic network depends on the choice of root, for this example nodes other than R1 were also tested. For example, when the node R2 was chosen, the log likelihood of the network was $-1,501.12$. Thus, it makes sense to choose as a root some vertex that maximizes the likelihood. In fact, from

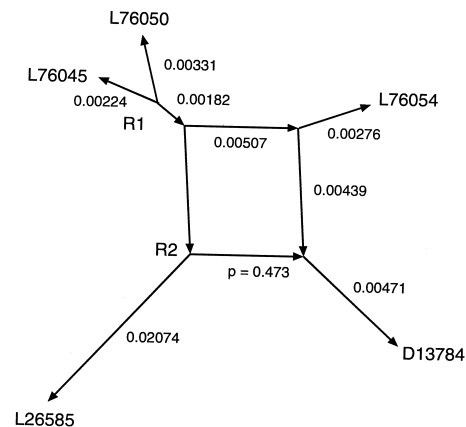


FIG. 6.—Network selected for the analysis of the HTLV data. Branch lengths and the prior p are maximum-likelihood estimates obtained using root R1 ($\log L = -1,500.32$).

an embedding of the investigated sequences in a larger network produced with SplitsTree, it was seen that node R1 was biologically preferable (data not shown).

Discussion

We have presented a likelihood approach to phylogenetic networks. The procedure is based on a directed graphical model for the evolution of sequences along a network and is applicable to sequences of any alphabet (nucleotides, amino acids). If the underlying evolutionary hypothesis is a tree, the approach of Felsenstein (1981) is recovered. This opens up a variety of possibilities for the statistical analysis of networks. Most importantly, assigning likelihoods to networks allows comparison with competing evolutionary tree/network hypotheses on a statistically sound basis. In addition, it provides a method for determining an optimal root for a given network and for estimating corresponding branch lengths and recombination parameters from the data using the maximum-likelihood principle. We have exemplified this using a viral data set.

In addition, parameters for the model of substitution and rate variation can be optimized on the basis of a network, and ancestral sequences at internal nodes can be inferred (Yang, Kumar, and Nei 1995). A series of statistical tests using the likelihood ratio (Huelsenbeck and Rannala 1997) or parametric bootstrapping (Goldman 1993) are applicable. The method could also be used to detect recombination events (Grassly and Holmes 1997). Finally, the structure of the network can be directly inferred from the data using, for example, procedures similar to that of learning general Bayesian networks (Buntine 1996; Krause 1998; Jordan 1999). Essentially, this amounts to devising search strategies similar to those employed for computing maximum-likelihood trees (Swofford et al. 1996; Larget and Simon 1999).

The current maximum-likelihood framework provides a unification of the theory of evolutionary trees with that of phylogenetic networks. Thus, networks no longer stand apart when a probabilistic treatment is necessary. Interpretation and understanding of phylogenetic networks is therefore greatly improved. However, to render analysis of large data sets possible, more work has to be done to improve the marginalization procedure and to explore suitable network search methods.

Acknowledgments

We thank Anne-Mieke Vandamme and Marco Salemi for organizing the stimulating Fifth European Workshop on Virus Evolution and Molecular Epidemiology, Leuven, 1999, where this work began. We also thank Daniel Moynet for providing us with the HTLV data set, and Arndt von Haeseler for helpful discussion. K.S. was supported by a Helmholtz bioinformatics position provided by Hans-Werner Mewes, and V.M. was supported by the Swedish Natural Science Research Council (NFR).

LITERATURE CITED

- BANDELT, H.-J. 1994. Phylogenetic networks. *Verh. Naturwiss. Ver. Hambg.* **34**:51–71.
- BANDELT, H.-J., and A. W. M. DRESS. 1992. A canonical decomposition theory for metrics on a finite set. *Adv. Math.* **92**:47–105.
- . 1993. A relational approach to split decomposition. Pp. 123–131 in O. OPITZ, B. LAUSEN, and R. KLAR, eds. *Information and classification*, Springer, Berlin.
- BANDELT, H.-J., P. FORSTER, B. C. SYKES, and M. B. RICHARDS. 1995. Mitochondrial portraits of human populations using median networks. *Genetics* **141**:743–753.
- BUNEMAN, P. 1971. The recovery of trees from measures of dissimilarity. Pp. 387–395 in F. R. HODSON, D. G. KENDALL, and P. TAUTU, eds. *Mathematics in the archeological and historical sciences*. Edinburgh University Press, Edinburgh, Scotland.
- BUNTINE, W. 1996. A guide to the literature on learning probabilistic networks from data. *IEEE Trans. Knowl. Data Eng.* **8**:195–210.
- CHICKERING, D. M., and D. HECKERMAN. 1997. Efficient approximations for the marginal likelihood of Bayesian networks with hidden variables. *Machine Learning* **29**:181–212.
- DOOLITTLE, W. F. 1999. Phylogenetic classification and the universal tree. *Science* **284**:2124–2128.
- DRESS, A. W. M., D. H. HUSON, and V. MOULTON. 1996. Analyzing and visualizing sequence and distance data using SplitsTree. *Discrete Appl. Math.* **71**:95–109.
- DURBIN, R., S. EDDY, A. KROGH, and G. MITCHISON. 1998. *Biological sequence analysis: probabilistic models of proteins and nucleic acids*. Cambridge University Press, Cambridge, England.
- EDWARDS, A. W. F., and L. L. CAVALLI-SFORZA. 1964. Reconstruction of evolutionary trees. Pp. 67–76 in V. H. HEYWOOD and J. MCNEILL, eds. *Phenetic and phylogenetic classification*. Systematic Association, London.
- FELSENSTEIN, J. 1981. Evolutionary trees from DNA sequences: a maximum-likelihood approach. *J. Mol. Evol.* **17**:368–376.
- FELSENSTEIN, J., and G. A. CHURCHILL. 1996. A hidden Markov model approach to variation among sites in rate of evolution. *Mol. Biol. Evol.* **13**:93–104.
- GELFAND, A. E., and F. M. SMITH. 1990. Sampling-based approaches to calculating marginal densities. *J. Am. Stat. Assoc.* **85**:398–409.
- GOLDMAN, N. 1993. Statistical tests of models of DNA substitution. *J. Mol. Evol.* **36**:182–198.
- GRASSLY, N. C., and E. C. HOLMES. 1997. A likelihood method for the detection of selection and recombination using nucleotide sequences. *Mol. Biol. Evol.* **14**:239–247.
- HOLMES, E. C., M. WOROBAY, and A. RAMBAUT. 1999. Phylogenetic evidence for recombination in dengue virus. *Mol. Biol. Evol.* **16**:405–409.
- HUELSENBECK, J. P., and B. RANNALA. 1997. Phylogenetic methods come of age: testing hypotheses in a evolutionary context. *Science* **276**:227–232.
- HUSON, D. H. 1998. SplitsTree: analyzing and visualizing evolutionary data. *Bioinformatics* **14**:68–73.
- JENSEN, F. V. 1996. *Introduction to Bayesian networks*. UCL Press, London.
- JORDAN, M. I. ed. 1999. *Learning in graphical models*. MIT Press, Cambridge, Mass.
- KASHAB, R. L., and S. SUBAS. 1974. Statistical estimation of parameters in a phylogenetic tree using a dynamical model of the substitutional process. *J. Theor. Biol.* **47**:75–101.

- KISHINO, H., and M. HASEGAWA. 1989. Evaluation of the maximum-likelihood estimate of the evolutionary tree topologies from DNA sequence data, and the branching order in Hominoidea. *J. Mol. Evol.* **29**:170–179.
- KRAUSE, P. J. 1998. Learning probabilistic networks. *Knowl. Eng. Rev.* **13**:321–351.
- LAKE, J. A., R. JAIN, and M. C. RIVERA. 1999. Genomics—mix and match in the tree of life. *Science* **283**:2027–2028.
- LARGET, B., and D. L. SIMON. 1999. Markov chain Monte Carlo algorithms for the Bayesian analysis of phylogenetic trees. *Mol. Biol. Evol.* **16**:750–759.
- LAURITZEN, S. 1996. Graphical models. Oxford University Press, Oxford, England.
- LIÒ, P., and N. GOLDMAN. 1998. Models of molecular evolution and phylogeny. *Genome Res.* **8**:1233–1244.
- MARGUSH, T., and F. R. MCMORRIS. 1981. Consensus n-trees. *Bull. Math. Biol.* **43**:239–244.
- NEYMAN, J. 1971. Molecular studies of evolution: a source of novel statistical problems. Pp. 1–27 in S. S. GUPTA and J. YACKEL, eds. *Statistical decision theory and related topics*. Academic Press, New York.
- OOTA, H., N. SAITOU, T. MATSUSHITA, and S. UEDA. 1999. Molecular genetic analysis of remains of a 2,000-year-old human population in China—and its relevance for the origin of the modern Japanese population. *Am. J. Hum. Genet.* **64**:250–258.
- PEARL, J. 1988. Probabilistic reasoning in intelligent systems: networks of plausible inference. Morgan Kaufmann, San Mateo, California.
- RUSSELL, S., and P. NORVIG. 1995. Artificial intelligence: a modern approach. Prentice Hall, Upper Saddle River, N.J.
- STRIMMER, K., and A. VON HAESLER. 1996. Quartet-puzzling: a quartet maximum-likelihood method for reconstructing tree topologies. *Mol. Biol. Evol.* **13**:964–969.
- SWOFFORD, D. L., G. J. OLSEN, P. J. WADELL, and D. M. HILLIS. 1996. Phylogenetic inference. Pp. 407–514 in D. M. HILLIS, C. MORITZ, and B. K. MABLE, eds. *Systematic biology*. Sinauer, Sunderland, Mass.
- TAMURA, K., and M. NEI. 1993. Estimation of the number of nucleotide substitutions in the control region of mitochondrial DNA in humans and chimpanzees. *Mol. Biol. Evol.* **10**:512–526.
- TEMPLETON, A. R. 1983. Phylogenetic inference from restriction endonuclease cleavage site maps with particular reference to the evolution of humans and the apes. *Evolution* **37**:221–244.
- VON HAESLER, A., and G. A. CHURCHILL. 1993. Network models for sequence evolution. *J. Mol. Evol.* **37**:77–85.
- YANG, Z., S. KUMAR, and M. NEI. 1995. A new method of inference of ancestral nucleotide and amino acid sequences. *Genetics* **141**:1641–1650.

APPENDIX A

Computer Programs and Data Set

Calculation of the maximum-likelihood values for the example networks was done using PAL (Phyloge-

netic Analysis Library), a Java package distributed by K.S. from <http://users.ox.ac.uk/~strimmer/pal/>. PAL is a collection of Java classes for use in molecular phylogenetics. For more information, please visit the web page. The HTLV data set is available on request.

APPENDIX B

Mathematical Definitions

In this appendix, we provide some definitions concerning split systems and outer-planar graphs, following A. W. M. Dress and D. H. Huson (personal communication).

If X is a finite set of cardinality n , then a *split* $S = \{A, \bar{A} := X - A\}$ of X is simply a bipartition of X (into two complementary nonempty sets). Two splits $S_1 = \{A_1, \bar{A}_1\}$ and $S_2 = \{A_2, \bar{A}_2\}$ are *compatible* if at least one of the four intersections $A_1 \cap A_2$, $A_1 \cap \bar{A}_2$, $\bar{A}_1 \cap A_2$, $\bar{A}_1 \cap \bar{A}_2$ is nonempty; otherwise, they are *incompatible*. A collection of splits Σ is called “*circular*” if there is an ordering x_1, x_2, \dots, x_n of the elements of X such that for every split $S \in \Sigma$ there exist p, q , with $1 < p \leq q \leq n$ and $S = \{\{x_p, x_{p+1}, \dots, x_q\}, X - \{x_p, x_{p+1}, \dots, x_q\}\}$.

Let $G = (V, E)$ be a finite connected graph, let C denote a set of *colors*, and let $\kappa : E \rightarrow C$ be an arbitrary *edge coloring*. Call κ an *isometric coloring* if the number of colors occurring in *every* shortest path between two vertices $v, w \in V$ is equal to minimal number of edges in *any* path from v to w , for all $v, w \in V$. A pair (G, κ) consisting of a connected bipartite graph G and an isometric edge coloring κ is called a *split graph*.

Now, given a collection Σ of splits of X , a split graph $(G = (V, E), \kappa : E \rightarrow \Sigma)$, together with a node labeling $l : X \rightarrow V$ is said to *represent* Σ if, for every $S = \{A, \bar{A}\} \in \Sigma$, the deletion of the all edges of color S produces a graph consisting of precisely two components, one containing all vertices labeled with elements of A and the other containing all vertices labeled with elements of \bar{A} . A *planar* split graph G (i.e., one that may be drawn in the plane without crossing edges) that represents Σ is called *outer-labeled* if all labeled vertices of G are of degree one and incident with the unbounded face of G .

All of the above concepts are tied together by the following theorem.

THEOREM. Σ is circular if and only if there exists an outer-labeled planar split graph that represents Σ .

MIKE HENDY, reviewing editor

Accepted February 3, 2000