JOURNAL OF COMPUTATIONAL BIOLOGY
Volume 11, Number 1, 2004
© Mary Ann Liebert, Inc.
Pp. 195–212

From a Phylogenetic Tree to a Reticulated Network

VLADIMIR MAKARENKOV[1,2] and PIERRE LEGENDRE[3]

ABSTRACT

In many phylogenetic problems, assuming that species have evolved from a common ancestor by a simple branching process is unrealistic. Reticulate phylogenetic models, however, have been largely neglected because the concept of reticulate evolution have not been supported by using appropriate analytical tools and software. The reticulate model can adequately describe such complicated mechanisms as hybridization between species or lateral gene transfer in bacteria. In this paper, we describe a new algorithm for inferring reticulate phylogenies from evolutionary distances among species. The algorithm is capable of detecting contradictory signals encompassed in a phylogenetic tree and identifying possible reticulate events that may have occurred during evolution. The algorithm produces a reticulate phylogeny by gradually improving upon the initial solution provided by a phylogenetic tree model. The new algorithm is compared to the popular *SplitsGraph* method in a reanalysis of the evolution of photosynthetic organisms. A computer program to construct and visualize reticulate phylogenies, called *T-Rex* (Tree and Reticulogram Reconstruction), is available to researchers at the following URL: *www.fas.umontreal.ca/biol/casgrain/en/labo/t-rex*.

Key words: least-squares fitting, phylogenetic tree, reticulate evolution, reticulated network, reticulate phylogeny.

INTRODUCTION

Evolution of species has long been assumed to be a branching process that could be represented only by a tree topology. In such a topology, a species can solely be linked to its closest ancestor; direct interspecies relationships (connection branches) are not allowed. Such well-known evolutionary mechanisms as hybridization or allopolyploidy cannot, however, be appropriately represented by means of a tree topology. Reticulate patterns of relationships have been found in a number of phylogenetic situations (Legendre, 2000): in bacterial evolution, lateral gene transfer is the mechanism allowing bacteria to exchange genes across species (Doolittle, 1999; Sneath, 2000); in plant evolution, allopolyploidy leads to the appearance of new species encompassing the chromosome complements of the two parent species; reticulate evolution is also present in microevolution within species in sexually reproducing eukaryotes

[1] Département d'informatique, Université du Québec à Montréal, C.P. 8888, Succ. Centre-Ville, Montréal (Québec), Canada, H3C 3P8.

[2] Institute of Control Sciences, 65 Profsoyuznaya, Moscow 117806, Russia.

[3] Département de sciences biologiques, Université de Montréal, C.P. 6128, succ. Centre-ville, Montréal, Québec H3C 3J7, Canada.

(Smouse, 2000). According to McDade (1995), analytical tools enabling one to generate reticulate topologies that accurately depict hybrid history represent a wide-open field for research. When traditional cladistic/phylogenetic methods are applied to such situations, they may produce confusing results since they are constrained to generate only treelike patterns. Homoplasy is another source of confusion in the reconstruction of phylogenetic trees; it can be represented by supplementary branches added to phylogenetic trees. In their review on reticulate evolution, Posada and Crandall (2001) considered several definitions of netlike evolution, accompanied by proposals of how the biological procedures involved should be represented mathematically. Nakhleh *et al.* (2003) have recently reported a suite of useful techniques for studying the topological accuracy of methods for reconstructing phylogenetic networks.

Several tentative methods have been proposed for discovering reticulate evolution in nucleotide sequences. Among existing works, we can mention displays of compatibility (Sneath, Sackin, and Ambler, 1975), tests for clustering (Stephens, 1985), a randomization approach (Sawyer, 1989), and an extension of the parsimony method of phylogenetic reconstruction that allows recombination (Hein, 1993). Rieseberg and Morefield (1995) developed a computer program, RETICLAD, for the identification of hybrids based on the expectation that they would combine the characters of their parents. The latter program can test only reticulate events between terminal branches of a tree. Rieseberg and Ellstrand (1993) showed some examples in which the program appears to work well. The popular method of split decomposition enables the representation of data in the form of a splitsgraph revealing the conflicting signals contained in the data (Bandelt and Dress, 1992a, 1992b; Bandelt, 1995). In a splitsgraph, a pair of nodes may be linked by a set of parallel branches depicting alternative evolutionary hypotheses. Hallet and Lagergren (2001) showed how lateral gene transfer events can be detected by comparing differences between species and gene trees. Bryant and Moulton (2002) introduced a network-inferring method, NeighborNet, allowing the reconstruction of planar phylogenetic networks. Each of these methods has features that make them useful for the analysis of particular types of data, and they all have a role to play in detecting and describing reticulate evolution.

In this article, we continue the development of a new method for detecting reticulate events in evolutionary data, which was first described in Legendre and Makarenkov (2002). We present a new algorithm for inferring reticulate phylogenies from evolutionary distances computed among species. This algorithm uses the topology of a phylogenetic tree as its supporting structure, from which a reticulated network is developed. We explore how new branches representing reticulate events can be added to a phylogenetic tree, transforming it into a reticulate phylogeny. The addition of each reticulation branch is done optimally using a least-squares criterion. The ins and outs of the new algorithm are shown by investigating the evolution of photosynthetic organisms. Analyzing the inferred reticulate phylogeny, we compare the novel approach to the widely used split decomposition technique (Bandelt and Dress, 1992a). Possible improvements of the reticulation model are also discussed (Appendix B); they would make it possible to construct a general reticulate structure not depending on the topology of a supporting tree. The proposed algorithm can also be applied to detect contradictory features in a given phylogenetic tree; the fewer the number of reticulation branches placed into a tree, the more credible the tree topology is.

MATERIALS AND METHODS

In this section, we describe the novel approach for reconstruction of *reticulated networks* representing the evolutionary relationships among a group of species (e.g., taxa). Mathematical definitions related to reticulated networks are given in Appendix A. Any reticulated network can be associated with a table of pairwise distances, called *reticulation distances*, between the nodes labeled by the names of the species; all other nodes of the network are intermediates: they represent unknown ancestors.

Buneman (1974) has shown that a distance matrix satisfying the four-point condition defines a unique phylogenetic tree. Reticulated networks are more general structures than phylogenetic trees; several different networks may be associated with the same *distance matrix*. For instance, the distance matrix given in Table 1 can be represented by a complete graph without intermediate nodes R_0 (Fig. 1*a*) or by the reticulated networks R_{11} and R_{12} (Figs. 1*b* and 1*c*) containing one intermediate node, or else by the reticulated networks R_{21} (Fig. 1*d*) or R_{22} (Fig. 1*e*) comprising two intermediate nodes. The nonuniqueness of reticulated networks suggests that a strong assumption about a possible reticulate topology should be made before starting the inference process.

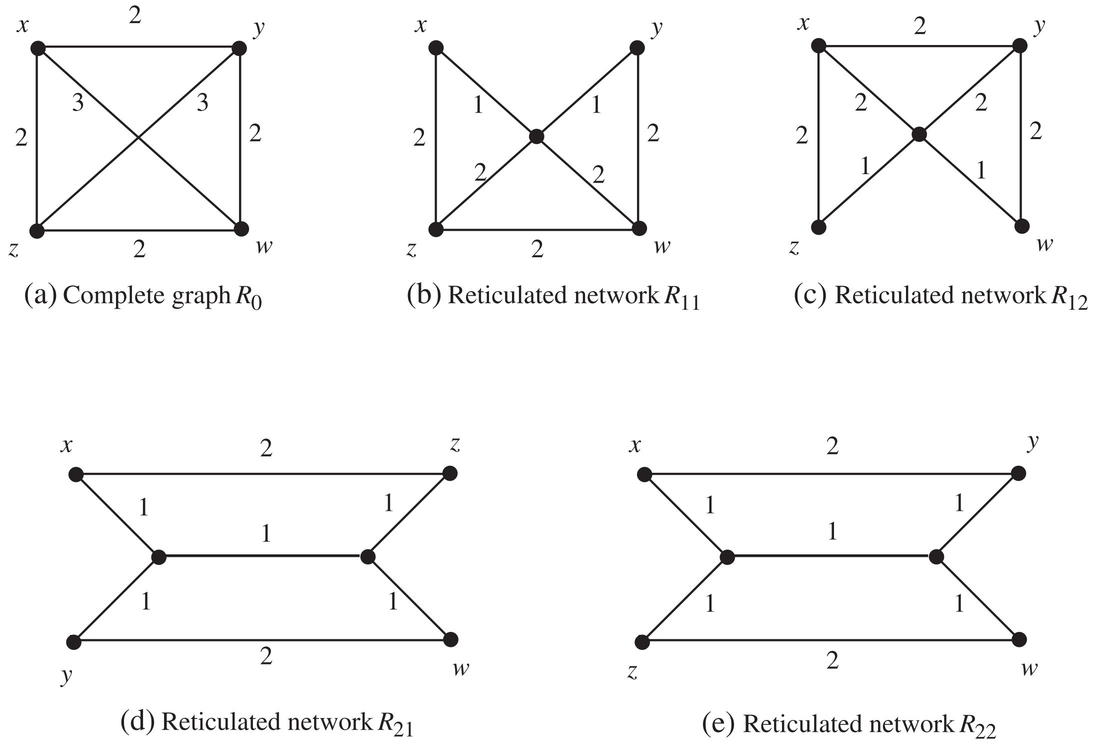
TABLE 1. DISTANCE MATRIX (RETICULATION DISTANCES) d FOR A SET OF TAXA x, y, z AND w

	x	y	z	w
x	0	2	2	3
y	2	0	3	2
z	2	3	0	2
w	3	2	2	0

In this study, we are using a phylogenetic tree topology as the basic structure for reconstructing a reticulated network. There are at least two justifications for this approach. First, in many evolutionary instances, a phylogenetic tree is already adequate to represent the evolution of a group of species, and in many cases, the number of reticulation events is small compared to the number of evolutionary events represented by the branches of a classical phylogenetic tree. Second, there exist a number of efficient and well-studied methods for inferring phylogenetic trees from distance data; see, for example, Saitou and Nei (1987), Gascuel (1997a), Felsenstein (1997), or Makarenkov and Leclerc (1999). These methods utilize different optimization criteria and should be applied whenever these criteria correspond to the assumptions made about the data at hand.

Algorithm for inferring a reticulated network

This section describes an algorithm for inferring a *connected* and *undirected reticulated network* (see Appendix A) from a given distance matrix. We propose the following approach to build a network from a matrix of evolutionary distances among observed taxa: first, infer a phylogenetic tree from a distance matrix using one of the existing tree fitting methods; supplementary branches, called *reticulation branches*, are then added to the tree structure, one at a time, each one minimizing a least-squares or a weighted



(a) Complete graph R_0

(b) Reticulated network R_{11}

(c) Reticulated network R_{12}

(d) Reticulated network R_{21}

(e) Reticulated network R_{22}

FIG. 1. A complete graph R_0 (**a**) and two reticulated networks, R_{11} (**b**) and R_{12} (**c**), with one intermediate node, as well as two possible reticulated networks, R_{21} (**d**) and R_{22} (**e**), with two intermediate nodes, associated with the distance matrix in Table 1.

least-squares loss function. The addition of reticulation branches stops when the minimum of a special *goodness-of-fit function* is reached. This function takes into account the value of the least-squares criterion as well as the total number of branches of the reticulated network under construction. Because in our study the reconstruction technique is based on the least-squares loss function, it is reasonable to consider an initial phylogenetic tree whose branch lengths have been fitted to the given distances by least squares. For an overview of least-squares fitting techniques, see Barthélemy and Guénoche (1991) or Bryant and Wadell (1998).

Let δ be a distance function on the set X of n taxa, and T a phylogenetic tree inferred from δ by means of an appropriate tree fitting method. Note that any given phylogenetic tree can be transformed into a *binary tree*, whose internal nodes are all of degree 3, by adding links of length zero where necessary. When this is done, a tree with n leaves has $n - 2$ internal nodes and $2n - 3$ branches. In this article, we consider binary phylogenetic trees as the foundation for the reticulated networks to be reconstructed. Thus, similarly to the binary trees, the reticulated networks considered in this study will comprise $2n - 2$ nodes. The original tree may be rooted or not; this does not matter when constructing undirected reticulated networks.

We will now explore how to place the first reticulation branch into a tree. To add a new branch to a phylogenetic tree, we will try out all possible pairs of nodes that are not already linked by a branch and select the one that reduces the value of the least-squares function the most. Let us consider a binary phylogenetic tree T inferred from a distance function δ and a pair of nodes x and y in T not linked by a branch (Fig. 2*a*). We look for an optimal value l, according to the least-squares loss function, for a potential new branch xy which may be added to the tree T, while keeping fixed the lengths of all preexisting tree branches (Fig. 2*b*).

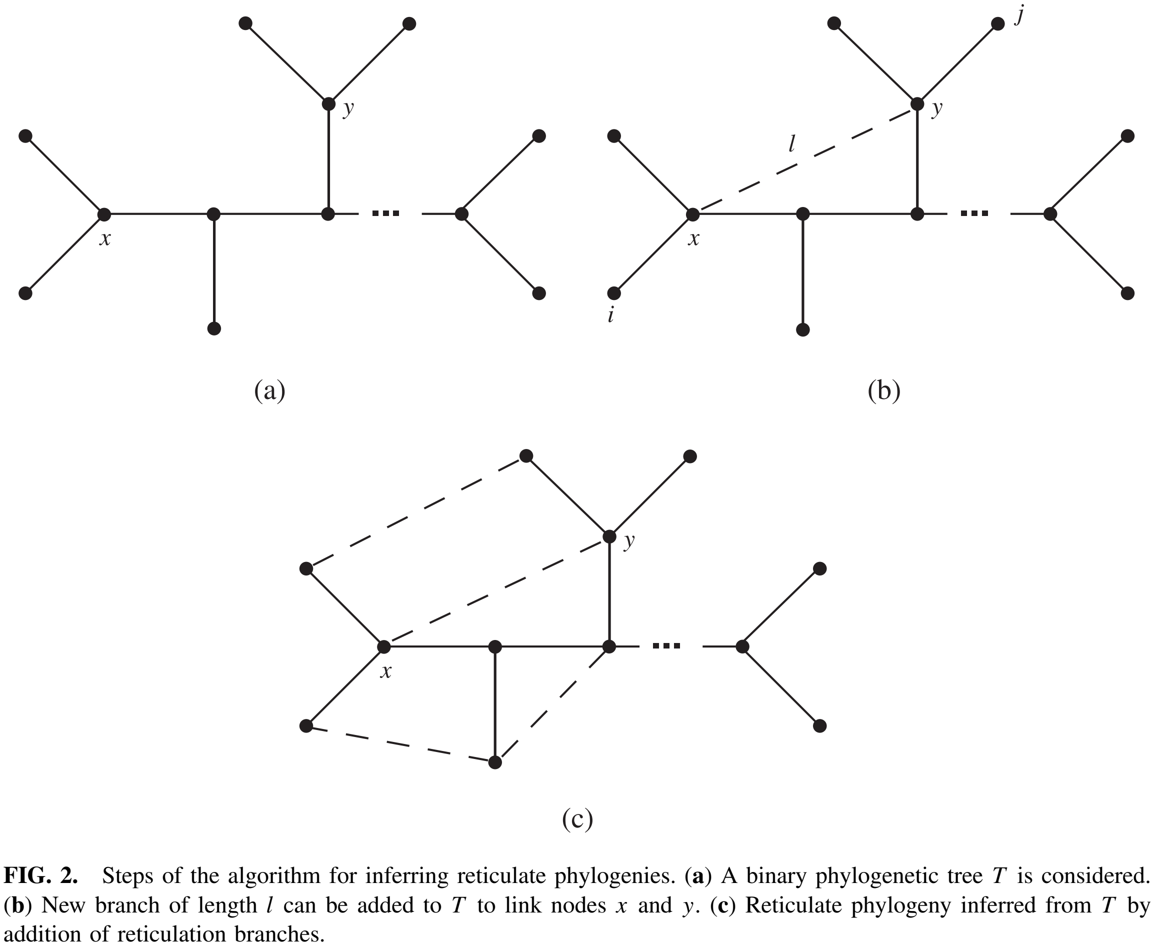


FIG. 2. Steps of the algorithm for inferring reticulate phylogenies. (**a**) A binary phylogenetic tree T is considered. (**b**) New branch of length l can be added to T to link nodes x and y. (**c**) Reticulate phylogeny inferred from T by addition of reticulation branches.

We will examine in greater detail how to determine the optimum value of the length of the first reticulation branch. First, we define a set $A(xy)$ representing the distances between pairs of taxa that are susceptible of changing if a new reticulation branch xy is added. Let d be a function of the distances in T between pairs of nodes. The set $A(xy)$ includes all pairs of taxa ij of X such that

$$Min\{d(i,x)+d(j,y); d(j,x)+d(i,y)\} < d(i,j). \tag{1}$$

For instance, if we add a new branch xy of length 0 to the tree, all distances between the pairs of taxa in $A(xy)$ will change their lengths. To find the optimal value l of a new potential branch xy, we have to subdivide $A(xy)$ into the m following subsets:

$$\begin{aligned} A_1 &= \{ij\} \text{ such that: } d(i,j) - Min\{d(i,x)+d(j,y); d(j,x)+d(i,y)\} \\ &= Min_{\{ij\in A(xy)\}}\{d(i,j) - Min\{d(i,x)+d(j,y); d(j,x)+d(i,y)\}\} = l_1; \\ A_k &= \{ij\} \text{ such that: } d(i,j) - Min\{d(i,x)+d(j,y); d(j,x)+d(i,y)\} \\ &= l_k > l_{k-1} \text{ (for } k = 2, \ldots, m-1), \\ A_m &= \{ij\} \text{ such that: } d(i,j) - Min\{d(i,x)+d(j,y); d(j,x)+d(i,y)\} \\ &= Max_{\{ij\in A(xy)\}}\{d(i,j) - Min\{d(i,x)+d(j,y); d(j,x)+d(i,y)\}\} = l_m = d(x,y) > l_{m-1}, \end{aligned}$$

where $A(xy) = \{A_1 \cup A_2 \cup \ldots \cup A_m\}$. The number of subsets m is the number of distinct values that the quantity $d(i,j) - \{d(i,x)+d(j,y); d(j,x)+d(i,y)\}$ can take over the set $A(xy)$.

The main reason for this subdivision is that each subset A_k has to be associated with an interval of possible length values l of the branch xy for which a particular optimization problem should be formulated. For each such optimization problem, a quadratic function has to be minimized, subject to a corresponding interval of length values of xy.

We will now show how the function to be minimized can be composed, for a fixed interval of branch length values, and how an optimal solution for this minimization problem can be found. Suppose that $l_k \le l \le l_{k+1}$, where $k = 0 \ldots m-1$. The constraint means that only the distances, i.e., the minimum-path-lengths $d(i,j)$, that are such that $ij \in \{A_m \cup A_{m-1} \cup \ldots \cup A_{k+1}\}$ will change lengths. We formulate the following problem to compute the optimal length value l of a potential new branch xy on the fixed interval $l_k \le l \le l_{k+1}$:

$$Q^*(xy,k) = \sum_{p=k+1}^{m} \sum_{ij\in A_p} (Min\{d(i,x)+d(j,y); d(j,x)+d(i,y)\} + l - \delta(i,j))^2 \to \text{min}. \tag{2}$$

Minimizing $Q^*(xy,k)$ minimizes the quadratic sum of differences between the values of the given evolutionary distance δ and the associated reticulation estimates. A nontrivial solution $l^*(xy,k)$ to this problem is the following:

$$l^*(xy,k) = \frac{\displaystyle\sum_{p=k+1}^{m} \sum_{ij\in Ap} (\delta(i,j) - Min\{d(i,x)+d(j,y); d(j,x)+d(i,y)\})}{\displaystyle\sum_{p=k+1}^{m} |Ap|}, \tag{3}$$

where the vertical bars denote the cardinality of the enclosed entity. If this quantity does not meet the constraint, the optimal solution $l^*(xy,k)$ has to be selected from the boundary values l_k and l_{k+1}. When $l^*(xy,k)$ and $Q^*(xy,k)$ have been found, the only remaining task is to compute the value of the least-squares objective function Q corresponding to this particular solution on the interval $l_k \le l \le l_{k+1}$ of length values of xy.

These computations have to be repeated over all intervals of branch lengths established for the given pair of nodes xy not linked by a branch. The global optimum value of the least-squares criterion Q, as well as the global optimum value of the branch length l over the set of defined intervals $l_k \leq l \leq l_{k+1}$, for $k = 0, \ldots, m-1$, are obtained recursively. To obtain the optimum value of Q over the set of all possible new branches, these computations should be repeated for all pairs of tree nodes that are not linked by a branch. Once the first reticulation branch has been added to the reticulated network, the best second, third, and following reticulation branches may be placed into it in the same way (Fig. 2*c*).

This algorithm takes $O(kn^4)$ time for n taxa and k new reticulation branches, since there are $O(n^2)$ taxon pairs ij for each pair of nonlinked nodes xy and $O(n^2)$ node pairs xy. Since all values of the reticulation distance d corresponding to an obtained reticulated network can only decrease, the reticulation distance always provides a more parsimonious solution than the initial additive distance (e.g., pairwise distance between taxa in the phylogenetic tree) from which it is derived. In Appendix B, we examine possible extensions of the reticulation model: reestimating all branch lengths, substituting a branch for another, or adding one or two intermediate nodes at once to the network.

Weighted least-squares criterion

Reticulation branches can also be added to the network according to a *weighted least-squares criterion* of the following form:

$$Q = \sum_{i \in X} \sum_{j \in X} w(i, j)[d(i, j) - \delta(i, j)]^2 \rightarrow \min. \tag{4}$$

The function $w(i, j)$ is applied to the separation of taxa i and j.

The weighted least-squares criterion may be useful in a number of evolutionary contexts. If some entries of the distance matrix are missing or uncertain, one can express this information through weighted least-squares by assigning low weights to the uncertain entries. If some values in the distance matrix are missing, such unknown data could be handled by setting the associated weights to zero. In the case of vicariance or other spatially constrained forms of phylogenetic problems, one can use binary weights to specify the groups of taxa among which reticulation branches are permitted, excluding the spatially separated pairs (see Example 1 in Legendre and Makarenkov [2002]). For an overview of applications of the least-squares and weighted least-squares criteria in the field of phylogenetics, readers are referred to Swofford and Olsen (1996) or Li (1998). There exist a number of efficient methods for inferring phylogenetic trees using weighted least squares: Felsenstein (1997) described how this kind of optimization is performed in the PHYLIP package; see also the papers by Makarenkov and Leclerc (1999) and Gascuel (2000) explaining how to reconstruct phylogenetic trees under different weighting conditions; on the other hand, the paper by Bryant and Wadell (1998) discusses how to compute optimal branch lengths for a tree with fixed topology in the weighted case. However, no important developments have taken place for the reconstruction of reticulate phylogenies using this important criterion.

The algorithm described earlier can easily be extended to the case of weighted least squares (Equation 4). The main difference compared to the unweighted case arises when the objective function is written over a fixed interval of length values ($l_k \leq l \leq l_{k+1}$, where $k = 0 \ldots n-1$) of the potential branch xy. In the weighted case, the function to be minimized is the following:

$$Q^* = \sum_{p=k+1}^{n} \sum_{ij \in A_p} w(i, j)[Min\{d(i, x) + d(j, y); d(j, x) + d(i, y)\} + l - \delta(i, j)]^2, \tag{5}$$

subject to $l_k \leq l < l_{k+1}$. A nontrivial solution $l^*(xy, k)$ for this minimization problem is as follows:

$$l^*(xy, k) = \frac{\displaystyle\sum_{p=k+1}^{m} \sum_{ij \in Ap} w(i, j)[\delta(i, j) - Min\{d(i, x) + d(j, y); d(j, x) + d(i, y)\}]}{\displaystyle\sum_{p=k+1}^{m} \sum_{ij \in Ap} w(i, j)}. \tag{6}$$

STOPPING RULES FOR ADDITION OF RETICULATION BRANCHES

A reticulated network comprises more branches and thus uses more parameters than a phylogenetic tree. As in all statistical models, more parameters mean better fit but fewer degrees of freedom and a loss of simplicity. A special cost criterion should be used to estimate how many reticulation branches have to be added to a network. We are proposing four possible goodness-of-fit criteria allowing one to determine when to stop adding branches to a reticulated network. All criteria take into account the least-squares objective function as well as the number of network parameters. When the exact number of reticulation branches is unknown in advance, as it is often the case in evolutionary problems, one can stop the addition of new branches when the minimum of the selected criterion is reached.

The total number of nodes in an unrooted binary phylogenetic tree with n leaves is $2n-2$. Therefore, the maximum number of branches one might place in a reticulated network, inferred from a binary phylogenetic tree with n leaves, is $(2n-2)(2n-3)/2$. However, any metric distance can be represented by a complete graph with $n(n-1)/2$ branches. Thus, any of these two limits $(2n-2)(2n-3)/2$ or $n(n-1)/2$ can be considered as the maximum possible number of branches in a reticulated network. If the latter limit is considered, the number of degrees of freedom of a reticulated network with N branches can be defined as $n(n-1)/2-N$.

It would be reasonable to consider a penalty function opposing the loss in degrees of freedom to the gain in fit. The first goodness-of-fit function that we consider is the following:

$$Q_1 = \frac{\sqrt{\sum_{i\in X}\sum_{j\in X}(d(i,j)-\delta(i,j))^2}}{n(n-1)/2-N} = \frac{\sqrt{Q}}{n(n-1)/2-N}. \tag{7}$$

The numerator of this function is the square root of the sum of quadratic differences between the values of the given distance δ and the corresponding reticulation estimates d. Interestingly, as was confirmed by a simulation study (see Legendre and Makarenkov, 2002), the function Q_1 usually has only one minimum over the interval $[2n-3, n(n-1)/2[$ of possible values of N. This minimum can define a stopping rule for addition of new branches to the reticulate phylogeny.

The least-squares function itself may also be used as an appropriate numerator for the goodness-of-fit measure. Thus, one can consider a slightly modified criterion, denoted Q_2, which usually adds more reticulation branches to the network than criterion Q_1. The stopping rule Q_2 was used in the application section below:

$$Q_2 = \frac{\sum_{i\in X}\sum_{j\in X}(d(i,j)-\delta(i,j))^2}{n(n-1)/2-N} = \frac{Q}{n(n-1)/2-N}. \tag{8}$$

One can also consider the Akaike information criterion (AIC) which is a useful and well-known statistic for model identification and evaluation (Akaike, 1987). A model with a minimum value of AIC may be chosen to be the best-fitting solution among several competing models. In our algorithm, the Akaike rule would select the model that minimizes the following quantity:

$$\mathrm{AIC} = \frac{Q}{(2n-2)(2n-3)/2-2N}. \tag{9}$$

Finally, another popular statistical estimator, the minimum description length (MDL) criterion introduced by Rissanen (1978), can also be used as a stopping rule in our algorithm. The MDL criterion, which is closely related to AIC statistics, can be computed as follows:

$$\mathrm{MDL} = \frac{Q}{(2n-2)(2n-3)/2-N\log(N)}. \tag{10}$$

MONTE CARLO STUDY

A Monte Carlo study was conducted to test the ability of the new method to cope with noisy phylogenetic data. To supplement the simulation study reported in Legendre and Makarenkov (2002), we will examine here how the new method reacts to the different kinds of noise condition affecting evolutionary data. All results presented below were obtained by simulating 1,000 random phylogenetic trees. In each case, a true phylogeny, denoted as T, was generated using the random phylogeny generation procedure described by Kuhner and Felsenstein (1994). The tree topologies were simulated by an iterative process in which, at each iteration, one of the k tree branches (where $k = 1, \ldots, 2n - 5$, for trees with n leaves) was chosen at random to be the one that splits. The lengths of the tree branches were drawn at random from an exponential distribution with expectation $1/(2n - 3)$. Then, following Guindon and Gascuel (2002), we added noise in the form of deviations from the molecular clock hypothesis. Every branch length of T was multiplied by $1 + ax$, where x was a value drawn at random from a standard exponential distribution ($P(x > k) = \exp(-k)$). The constant a was a tuning factor meant to adjust the deviation intensity from the molecular clock hypothesis; as in Guindon and Gascuel (2002), a was set at 0.8. The trees generated by this process were expected to have $O(\log(n))$ depth. The additive distance matrix $\mathbf{D}$ associated with the true tree T was then computed and normalized to have unit variance. The source code of our tree generation program, written in C++, is available to researchers at *www.info.uqam.ca/~makarenv/tree_generation.cpp*.

In this study, we show how the new method behaved in the situation when the observed data corresponded to a phylogenetic tree affected by different amounts of noise. Normally distributed random errors with mean zero and variances σ^2 varying from 0 to 0.3 were added to $\mathbf{D}$ to obtain replicates of the distance matrix Δ. In the rare cases where a negative value arose in Δ, it was replaced by the constant 0.01. The simulations were carried out with phylogenies with $n = 24$ leaves. The results, presented in Figs. 3*a*, 3*b*, and 3*c*, are averages for 1,000 different distance matrices. Figure 3*a* shows the neighbor-joining (NJ) (Saitou and Nei, 1987) topology recovery as a function of the amount of noise. Mean values of the Robinson and Foulds (RF) topological distance (Robinson and Foulds, 1981) between a true tree and a tree derived by NJ are shown; the RF distance was normalized by its greatest possible value, which is $2n - 6$ for a binary phylogenetic tree with n leaves. Nowadays, NJ is arguably the most popular method for constructing phylogenies from distance data. For some time, the success of NJ was inexplicable for computational biologists, due to the lack of approximation bounds (Bryant and Moulton, 2002). One of the first bounds was found by Atteson (1999), who showed that NJ would be able to return the true phylogeny given that the observed distance is sufficiently close to the true evolutionary distance. On the other hand, Gascuel (1997a,b) proved that the branch length estimation and distance matrix reduction formulae in NJ provide low variance estimators. In the paper describing the BioNJ method, Gascuel (1997a) showed how to improve NJ accuracy by incorporating minimum variance optimization in the NJ reduction formula. While observing the behavior of the NJ curve in Fig. 3*a*, one will note that, as expected, the performance of NJ decreases with increasing noise.

Figures 3*b* and 3*c* show the asymptotic behavior of the algorithm described in this paper in the situation where the original data correspond to a phylogenetic tree. No reticulation branches at all should be added to a tree when the generated distance matrix satisfies the four-point condition and, thus, perfectly corresponds to a tree topology. However, as in the type I error of statistical tests, we can expect some reticulation branches to be formed in the presence of noisy data. True phylogenies with 24 leaves were randomly generated and biased by noise as described above; then, phylogenetic trees were inferred from the noisy patristic distances by NJ. Following this, the new algorithm for reconstructing reticulate phylogenies was applied. We computed how many reticulation branches should have been added to the NJ phylogenies to provide the preselected amount of improvement in fit, varying from 1% to 25%, with respect to the least-squares coefficient Q obtained for the NJ phylogenies. In Fig. 3*b*, type I error was graphed against a fixed amount of improvement in fit (graphs for 10%, 15%, and 25% improvement are shown) for σ^2 varying from 0 to 0.3. In Fig. 3*c*, the variance of the noise σ^2 was fixed (graphs with $\sigma^2 = 0.1$, 0.2, and 0.3 are shown), and the amount of improvement in fit varied from 0% to 25%.

The following observations can be made after analyzing the graphs in Figs. 3*b* and 3*c*. The simulations show that no reticulation branches were added by the new algorithm when analyzing error-free data ($\sigma^2 = 0$). In the case of trees affected by noise, the number of reticulation branches necessary to produce a preselected gain in fit increased with increasing noise. First, the results suggest that the new method

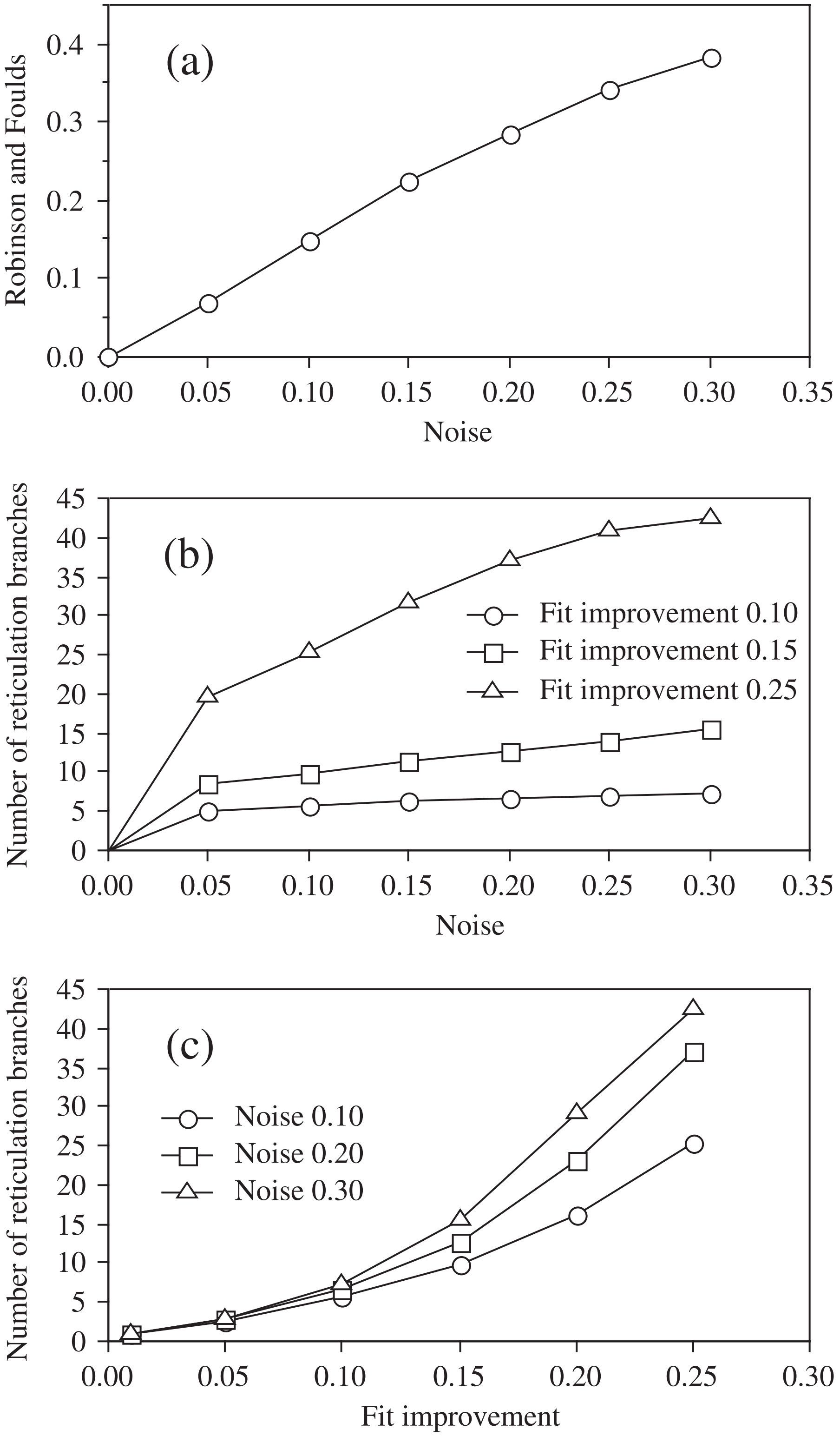


FIG. 3. Mean simulation results for 1,000 trees with 24 leaves. (**a**) Mean values of the normalized Robinson and Foulds topological distance between the true tree and the tree derived by NJ, as a function of the amount of noise. Smaller distances correspond to better recovery. (**b**) Fixed improvement in fit, varying noise: mean number of reticulation branches needed to obtain fixed improvements in fit of 10%, 15%, and 25% measured using the least-squares coefficient Q of the NJ tree. (**c**) Fixed noise, varying improvement in fit: mean number of reticulation branches as a function of improvement in fit for trees with fixed variance of the noise σ^2 of 0.1, 0.2, and 0.3.

is able to recognize a true phylogenetic tree by not adding any reticulation branch to it. Second, they indicate that when the new method is applied to analyze unreticulated but noisy data, it is likely to produce reticulation branches that will represent *contradictory features* existing in randomly generated or noise-biased data set. Third, the following trend can be observed: the more noise was added to the data set, i.e., the closer the distance matrix was to a completely random distance matrix, the greater number of *conflicting signals* were detected and, as a consequence, the greater number of reticulation branches were added to a phylogenetic tree to attain a preselected value of improvement in fit.

EVOLUTION OF PHOTOSYNTHETIC ORGANISMS

In this section, we show how the new algorithm for inferring reticulate phylogenies may help to examine the evolution of photosynthetic organisms. We provide a comparison of our technique with the popular splits-graph method, which is also meant to detect and represent contradictory features in evolutionary data. Some common features and differences between the two approaches are discussed.

Because a set of real evolutionary data may contain a number of conflicting signals, evolutionary events cannot always be modeled as a treelike process. To address this problem, Bandelt and Dress (1992a) designed the method of split decomposition allowing one to transform evolutionary data into a sum of weakly compatible splits. There exist a number of algorithms for carrying out split decomposition (see Bandelt and Dress [1992b] or Huson [1998]). In this study, we used the second version of the SplitsTree program by Huson (1998).

Let us recall some basic definitions related to splits-graphs. Let X be a set of taxa. A split $S = \{B, B'\}$ is defined as a partition of X into two nonempty sets B and B' such that $B \cup B' = X$. For instance, any branch in a phylogenetic tree introduces a split consisting of all the taxa found on one side (set B) and on the other (set B') of this branch. A set S of splits is called *weakly compatible* if, for any three splits S_1, S_2, and S_3 from S and all $B_i \in S_i$ $(i = 1, 2, 3)$, at least one of the four intersections

$$B_1 \cap B_2 \cap B_3,\ B_1 \cap B_2' \cap B_3',\ B_1' \cap B_2 \cap B_3',\ \text{or}\ B_1' \cap B_2' \cap B_3$$

is empty (see Bandelt and Dress, 1992a, b). A *splits-graph* representing a weakly compatible split system S is a graph $G(S) = (V, E)$ whose nodes $v \in V$ are labeled by the set of taxa in X and whose branches $e \in E$ are straight line segments that represent the splits in S; see Fig. 4*a*. In such a graph, each split $\{B, B'\}$ in S is represented by a group of parallel branches of equal lengths, so that deleting all branches in such a group splits the graph into exactly two parts, one containing all nodes labeled by the taxa in B and the other containing all nodes labeled by the taxa in B'.

Table 2 contains the pairwise distances among eight species of photosynthetic organisms. The original sequence data, obtained by Lockhart *et al.* (1993), consist of 920 bases from the 16S rRNA of the chloroplasts of algae, liverworth, and higher plants, and also of a cyanobacterium. The data are available among the examples distributed with the SplitsTree program. The distance matrix was calculated by Huson (1998) using the log-determinant distance (LogDet) introduced by Steel (1994) and Lockhart *et al.* (1994).

The derived splits-tree path-length distances, obtained from the SplitsTree program, are reported in Table 3; the splitsgraph in shown in Fig. 4*a*. It illustrates evolutionary conflicts existing between the cyanobacterium and the chloroplasts of *Euglena* and the chrysophyte. The correct phylogenetic split should put *Euglena* (which contains chlorophyll a and b) together with the other chl-a/b containing species: rice, tobacco, liverworth, *Chlamydomonas*, *Chlorella*. Chrysophytes (also called chromophytes) contain chlorophyll a and c, while cyanobacteria contain only chlorophyll a. The main reason for the observed conflicting signal is that the rRNA chloroplast sequences in *Euglena* and the chrysophyte have, probably independently, acquired similar base compositions (convergence); see the discussion in Huson (1998) and in the references therein.

In a splits-graph, conflicting information is represented by parallel branches that can be removed to create a split forming subsets of taxa. (a) The largest set of parallel branches has lengths 0.0142; by cutting them, one separates a group containing all the species with chlorophyll a and b from the chrysophyte and cyanobacterium. (b) The next split to consider has length 0.0087. It would isolate *Euglena* and the chrysophyte from all the other species. We don't know how to interpret this split. (c) The third split, of

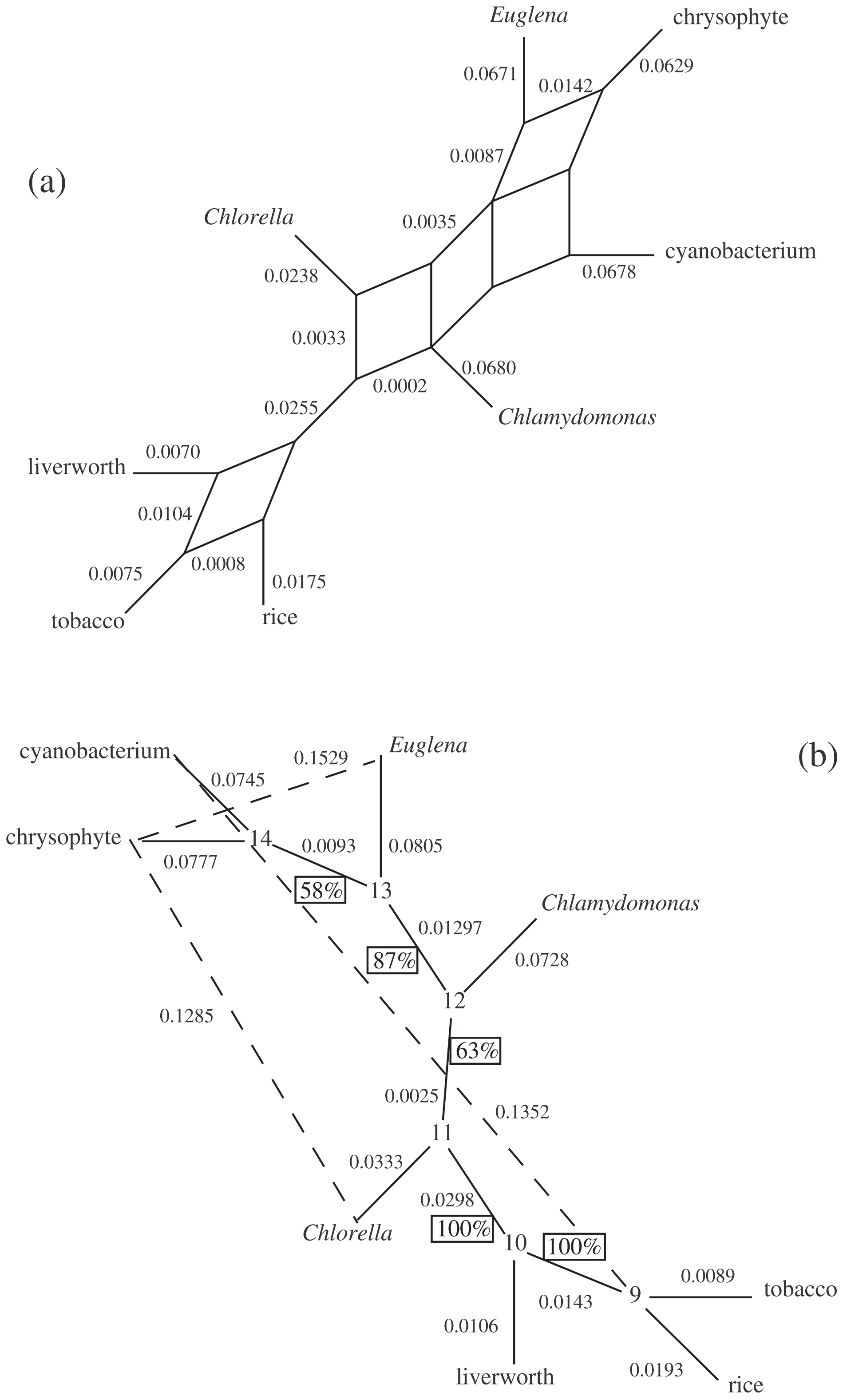


FIG. 4. (**a**) Splits-graph associated with the distances in Table 3. Each band of parallel branches indicates a possible split. (**b**) Reticulate phylogeny associated with the distances in Table 4. The reticulated network was constructed by adding three branches (dashed lines) to the NJ (neighbor-joining) phylogenetic tree (full lines). For the sake of a better representation, all tree branches are drawn equal. The numbers on the branches represent their lengths. Boxes: bootstrap support values for the clades on the NJ tree.

TABLE 2. PAIRWISE DISTANCES AMONG SPECIES OF PHOTOSYNTHETIC ORGANISMS OBTAINED FROM THE SPLITSTREE PROGRAM USING THE LOGDET DISTANCE OVER A SET OF CHLOROPLAST 16S rRNA SEQUENCES[a] (SEE HUSON, 1998)

Tobacco	0.0000							
Rice	0.0282	0.0000						
Liverworth	0.0324	0.0455	0.0000					
Chlamydomonas	0.1290	0.1386	0.1107	0.0000				
Chlorella	0.0875	0.0991	0.0699	0.1132	0.0000			
Euglena	0.1524	0.1621	0.1366	0.1621	0.1271	0.0000		
Cyanobacterium	0.1435	0.1551	0.1400	0.1657	0.1372	0.1791	0.0000	
Chrysophyte	0.1601	0.1670	0.1508	0.1807	0.1285	0.1529	0.1521	0.0000

[a] Liverworth: *Marchantia sp.*; cyanobacterium: *Anacystis nidulans*; chrysophyte (or chromophyte): *Olisthodiscus luteus*.

TABLE 3. PATH-LENGTH DISTANCES AMONG SPECIES OF PHOTOSYNTHETIC ORGANISMS IN THE SPLITS-GRAPH IN FIG. 4a OBTAINED AFTER USING, AS INPUT, THE PAIRWISE DISTANCES IN TABLE 2 (SPLITSTREE PROGRAM)

Tobacco	0.0000							
Rice	0.0258	0.0000						
Liverworth	0.0248	0.0357	0.0000					
Chlamydomonas	0.1124	0.1215	0.1014	0.0000				
Chlorella	0.0713	0.0804	0.0604	0.0920	0.0000			
Euglena	0.1270	0.1361	0.1161	0.1506	0.1033	0.0000		
Cyanobacterium	0.1299	0.1390	0.1190	0.1535	0.1128	0.1611	0.0000	
Chrysophyte	0.1370	0.1461	0.1261	0.1606	0.1133	0.1442	0.1427	0.0000

TABLE 4. PATH-LENGTH DISTANCES AMONG SPECIES OF PHOTOSYNTHETIC ORGANISMS IN THE RETICULATED PHYLOGENY IN FIG. 4b, OBTAINED AFTER USING, AS INPUT, THE PAIRWISE DISTANCES IN TABLE 2 (T-REX PROGRAM)

Tobacco	0.0000							
Rice	0.0283	0.0000						
Liverworth	0.0337	0.0441	0.0000					
Chlamydomonas	0.1283	0.1387	0.1157	0.0000				
Chlorella	0.0862	0.0966	0.0736	0.1086	0.0000			
Euglena	0.1490	0.1594	0.1364	0.1663	0.1293	0.0000		
Cyanobacterium	0.1441	0.1545	0.1396	0.1695	0.1325	0.1643	0.0000	
Chrysophyte	0.1554	0.1658	0.1428	0.1727	0.1285	0.1529	0.1521	0.0000

length 0.0035, would isolate *Euglena*, the chrysophyte and the cyanobacterium from the other species (two chlorophytes, liverworth, and the higher plants). We don't know how to interpret this split either. (d) The fourth split, of length 0.0033, is more interesting: it puts together three species (*Chlorella*, *Euglena*, and the chrysophyte) that are known to be facultative heterotrophs; i.e., they have the capacity to use organic substrates while growing in complete darkness (Stevenson *et al.*, 1996, Section 10.I.A).

We also carried out the analysis of distance data in Table 2 using the algorithm described in this paper. First, a phylogenetic tree was inferred from Table 2 by means of the method of weights MW, providing optimal least-squares estimates for the tree branches (see Makarenkov and Leclerc, 1999). Bootstrap support values were computed using NJ; the topology of the NJ tree was identical to that of the MW tree. The least-squares coefficient Q obtained by this approximation of the original data was 0.000917. The inferred phylogenetic tree (full lines) is shown in Fig. 4*b*, together with the three reticulation branches added by the new algorithm (dashed lines). The path-length distances between species in the reticulation structure are reported in Table 4. Note that the phylogenetic tree put *Euglena* together with the other species containing chlorophyll *a* and *b* (rice, tobacco, liverworth, *Chlamydomonas*, *Chlorella*).

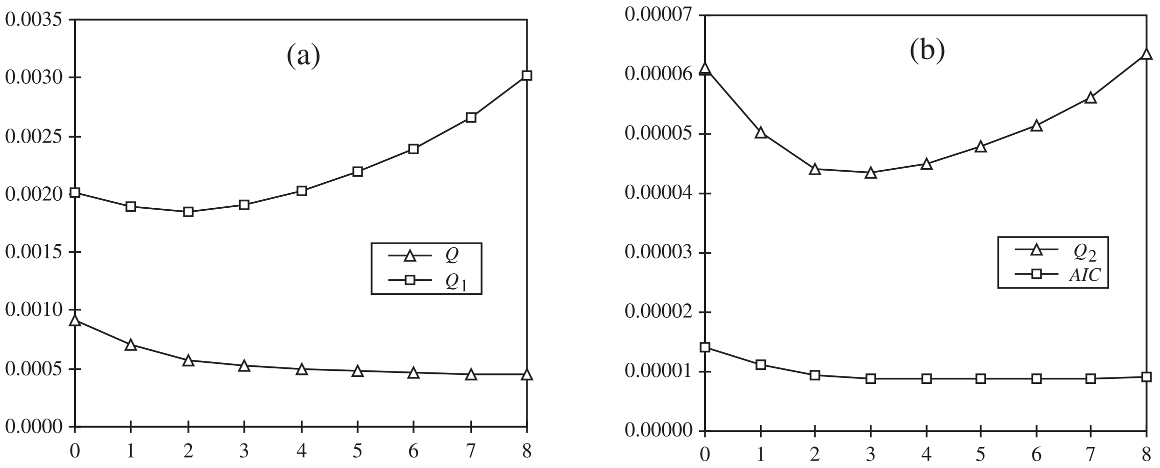


FIG. 5. Behavior of (**a**) the least-squares function Q (open triangles) and the goodness-of-fit criterion Q_1 (open squares), and (**b**) of the goodness-of-fit criteria Q_2 (open triangles) and AIC (open squares) for the first eight iterations of the algorithm for inferring reticulated networks applied to the pairwise distances among species of photosynthetic organisms from Table 2. Abscissa: number of iterations of the algorithm. Zero corresponds to the phylogenetic tree before reticulation branches were added. The minima of Q_1, Q_2, and AIC were reached at iterations 2, 3, and 4, respectively.

When building a reticulated network, the stopping rules Q_1, Q_2, and AIC discussed above suggested different solutions: the minimum value of Q_1 was found at the second iteration (Fig. 5*a*), the minimum of Q_2 at the third one (Fig. 5*b*), and the minimum of AIC at the fourth one (Fig. 5*b*). The solution with three reticulation branches, obtained using criterion Q_2, is shown in Fig. 4*b*. The first reticulation branch, linking *Euglena* to the chrysophyte, decreased the value of Q to 0.000704. The second one linked the cyanobacterium to node 9, which is the attachment point for the higher plants, tobacco, and rice; it decreased the value of Q to 0.000573. The third one, linking *Chlorella* to the chrysophyte, decreased the value of Q to 0.000522. The reticulation branches linking *Euglena*, *Chlorella*, and the chrysophyte delineate the group of facultative heterotrophs described above. On the other hand, the long reticulation branch (length $= 0.1352$) found between the cyanobacterium and the higher plants, which is an order of magnitude longer than the branches of the tree, is suggestive of the endosymbiosis hypothesis of Margulis (1981). According to this hypothesis, the cytoplasmic organelles (plastids) found in the cells of eukaryotes are thought to have once been free-living primitive bacteria that have become symbionts living inside the eukaryotic cells. Chloroplasts, in particular, could be derived from primitive cyanobacteria. Endosymbiosis is a form of lateral gene transfer that occurred in the deep phylogeny.

Let us now compare the numerical results provided by the SplitsTree program and the new algorithm for inferring reticulate phylogenies. The value of the least-squares coefficient Q computed for the splitsgraph in Fig. 4*a* is 0.008739. This result is not nearly as good as the value of $Q = 0.000522$ obtained for the reticulated network in Fig. 4*b*. As to the cophenetic correlation computed for the two structures in Figs. 4*a* and 4*b*, the reticulate phylogeny also compares advantageously to the splitsgraph with correlation coefficients equal to 0.995365 and 0.990795, respectively.

DISCUSSION

We have developed an algorithm to infer reticulate phylogenies from evolutionary distances among observed species. The new algorithm builds a reticulated network by adding supplementary branches to a phylogenetic tree. Any new branch added to a phylogenetic tree represents unresolved conflicting information contained in it. Two species or clades that are linked by a reticulation branch are more closely related to one another than in the phylogenetic tree representation that provided the initial fit for the evolutionary distances. The main challenge consists in giving plausible explanations for each of the extra relations represented by reticulation branches. These new branches should be interpreted differently under different evolutionary circumstances. First, we suggest that long reticulation branches linking nodes located far away from one another in the phylogenetic tree reveal incompatibilities of a tree structure with respect to the observed evolutionary distances. Two explanations are possible in this situation: first, the phylogenetic tree does not provide a good representation of the evolutionary distances; second, long reticulation branches may represent homoplasy among the observed species.

For data that are assumed to comprise neither reticulate relationships nor any homoplasy, reticulation analysis can be used to decide which phylogenetic tree, among a set of trees of nearly the same length, is the best one to represent the data: a tree containing fewer reticulation branches, especially the long ones, may be seen as the one embedding less conflicting signal. Special attention should be paid to short reticulation branches linking nodes located near one another. These branches may reflect either hybridization events that occurred between related species or their ancestors, or allopolyploidy if plant genetic distances are considered. The case of lateral gene transfer (LGT) seems to be the most complicated one because reticulation branches depicting gene exchange may be of any length. In this situation, investigation of the characters causing a reticulation might assist in the interpretation: if the responsible characters are contiguous in the nucleic acid sequence, LGT can be indicated.

We would like to give some recommendations to researchers who have access to sets of molecular sequences of certain species and want to test the data for presence of reticulate evolution. First, a matrix of evolutionary distances among the species has to be computed using an appropriate distance transformation. Among the most popular transformation functions are the Hamming, Kimura 3ST (Kimura, 1981), Jukes Cantor (Jukes and Cantor, 1969) and LogDet (Steel, 1994) transformations. Second, a phylogenetic tree has to be inferred from the distance matrix using a tree fitting algorithm. Third, the algorithm for reconstructing reticulate phylogenies can be applied, using one of the goodness-of-fit criteria as a stopping rule for addition of reticulation branches. We recommend to verify the solutions obtained using all four stopping rules (Q_1, Q_2, AIC, and MDL) discussed above and retain for further investigation the most interpretable reticulate phylogeny. In some situations, especially when the initial fit of the distance data provided by a phylogenetic tree is good, all four stopping rules may suggest adding the same number of reticulation branches to the phylogenetic tree. Interpretation of the reticulation branches should be done using the biological or evolutionary knowledge available about data at hand.

The algorithm for reconstructing reticulate phylogenies introduced in this paper has been included in the T-Rex (tree and reticulogram reconstruction) package (see Makarenkov, 2001). The T-Rex program, implemented for Windows and Macintosh platforms, is freely available for researchers at *www.fas.umontreal.ca/biol/casgrain/en/labo/t-rex*. The package also includes some popular phylogenetic tree fitting algorithms: ADDTREE by Sattath and Tversky (1977), neighbor-joining (NJ) by Saitou and Nei (1987), unweighted neighbor-joining (UNJ) by Gascuel (1997b), the method of weighted least-squares (MW) by Makarenkov and Leclerc (1999), circular order reconstruction by Makarenkov and Leclerc (1997, 2000), and others. T-Rex allows users to infer and visualize reticulate phylogenies by adding extra branches to phylogenetic trees obtained by the above-mentioned tree fitting algorithms.

APPENDIX A: FORMAL DESCRIPTION OF A RETICULATED NETWORK

This appendix provides definitions concerning reticulated networks. A reticulated network R is a weighted graph defined as a triplet (V, B, l), where V is a set of nodes (points or vertices), B is a set of branches (links), and l is a *function* of branch lengths assigning real nonnegative numbers to the branches. Each node i is either a taxon belonging to a set X or an intermediate node belonging to $V - X$. Thus, there are two different types of nodes in R.

A *path* p from node i to node j in R is a sequence of branches, $b_1b_2, b_2b_3, \ldots b_{k-1}b_k$, with $b_1 = i$ and $b_k = j$. The length of path p is given by the sum of the lengths of branches included in p and is denoted $l_p(i, j)$. A reticulated network is *connected* if, for every pair of nodes i and j, there exists at least one path from i to j. It is called *undirected* if there is no direction associated with the branches. Given a connected and undirected reticulated network R, the *minimum-path-length distance* between nodes i and j, denoted $d(i, j)$, is defined as in any weighted graph: $d(i, j) = min\{l_p(i, j) | p$ is a path from i to $j\}$.

A set of *reticulation distances* can be associated with the set of pairwise distances among the taxa in X. They are the minimum-path-length distances among taxa whose relationships are represented by a reticulated network.

APPENDIX B: EXTENSIONS OF THE RETICULATION MODEL

Here we discuss some possible modifications and improvements of the reticulation model presented in this paper. First, we consider the problem of reestimating the branch lengths of a reticulated network. Then, we examine the case where one of the branches is removed from the reticulated network and the case where one or two new intermediate nodes are added to it. Although these operations may make the reticulation model more complex and increase the time complexity of the inferring procedure, they may allow one to build a generic reticulate phylogeny which is totally independent of the basic phylogenetic tree.

In the algorithmic section, a stepwise optimization procedure designed to add a single reticulation branch at a time was described. It was intended to optimize the choice of the new branch as well as its length. Interestingly, the same calculations can be used to update all the other branch lengths. To reassess the length value of any branch in a reticulate phylogeny, one can use again Equations 2 and 3 assuming that the lengths of all the other branches are fixed. After a new reticulation branch has been added to a network, the polishing procedure can be carried out for branch number 1, then branch number 2, and so on, until all branch lengths are optimally reestimated. Then, one can return to the new reticulation branch to reassess its length for the second time, and so forth. The reestimation loop may be repeated several times to achieve the minimum value of the least-squares function for the reticulated network with a fixed topology. As this is usually the case, improvement in fit requires an increase in time complexity. If the reestimation procedure described above is incorporated into the algorithm, the time complexity of each iteration will increase up to $O(pmn^4)$, where m is the number of branches in the reticulated network and p is the number of loops performed over all branches.

Another operation which could improve the fitting precision consists in removing an existing branch and adding a new one; in other words, substituting a branch for another. All branches, including those of the original phylogenetic tree, could be candidates for removal. The only restriction for this operation is that the resulting network must not become disconnected. For a particular branch ab of length l_{ab} considered for removal from the reticulated network R, we have to find all pairs of taxa that will be affected by this deletion. This means that for any pair of taxa ij such that either $d(i,a) + l_{ab} + d(b,j) = d(i,j)$ or $d(j,a) + l_{ab} + d(b,i) = d(i,j)$, we have to recompute the value $d(i,j)$ under the condition that the branch ab is no longer in R. This operation can be followed by the branch addition operation. The pair of branches (removed and added) corresponding to the lowest value of the least-squares function Q may be selected for substitution. This operation may significantly redesign the topology of the reticulated network, which was initially based on a phylogenetic tree. Some branches of the original phylogenetic tree may no longer be part of the reticulated network. The time complexity of the removing–adding operation is $O(mn^4)$, where m is the number of branches in the reticulated network and n is the number of taxa. If only the branch removal operation (not cumulative with branch addition) is considered, we simply have to recompute the value of the goodness-of-fit criterion and make the decision about the potential branch deletion.

One may also want to introduce new intermediate nodes into the reticulated network. To deal with this issue, one has to consider a new optimization problem. Suppose that a new node y belonging to a new branch xy, which we are attempting to add to the reticulated network R, splits an existing branch zw of length l_{zw} into two parts yz and yw, as shown in Fig. 6*a*. One has to consider all pairs of taxa ij such that the associated distances $d(i,j)$ are susceptible of changing when the branch xy is added. To simplify the problem, one may assume that the branch zw is such that $l_1 + l_2 = l_{zw}$, where l_{zw} is a fixed length value. Similarly to the optimization problem for addition of one reticulation branch (Equation 2), a particular minimization problem can be formulated in the case of addition of a new branch with a new node. In this new problem, the minimization will also be done under constraint, but the optimization will involve three unknown branch lengths l, l_1, and l_2 instead of one.

To address the more complicated problem consisting of adding two intermediate nodes at once to the reticulated network (in Fig. 6*b*, two new nodes x and y and a new branch xy linking them are considered

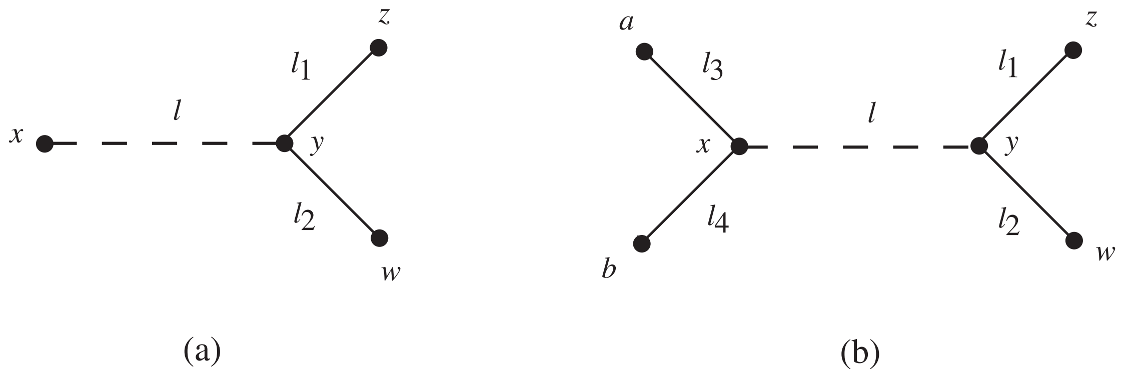


FIG. 6. (**a**) Incorporation of a new intermediate node y along with a new branch xy into a reticulated network. (**b**) Incorporation of two new intermediate nodes x and y along with a new branch xy into a reticulated network.

for addition), one could assume that the new branch xy splits the existing branches ab and zw in their middle. Thus, the only unknown variable is the length l of the new branch xy. One can then carry out the estimation procedure (Equations 2 and 3) to compute the optimal length l of xy while keeping all the other branch lengths fixed. This procedure may be followed by the polishing procedure for branch length reestimation, carried out over a whole network.

ACKNOWLEDGMENTS

The authors are grateful to the biologist and software developer Philippe Casgrain for his contribution to programming the T-Rex package, which makes the new method of reticulation analysis available to the scientific community, and to Dr. Yves Desdevises for help during the phylogenetic analyses.

REFERENCES

Akaike, H. 1987. Factor analysis and AIC. *Psychometrika* 52, 317–332.

Atesson, K. 1999. The performance of neighbor-joining methods of phylogenetic reconstruction. *Algorithmica* 25, 251–278.

Bandelt, H.-J. 1995. Combination of data in phylogenetic analysis. *Plant Systematics and Evolution Supplementum* 9, 355–361.

Bandelt, H.-J., and Dress, A.W.M. 1992a. Split decomposition: A new and useful approach to phylogenetic analysis of distance data. *Molecular Phylogenetics and Evolution* 1, 242–252.

Bandelt, H.-J., and Dress, A.W.M. 1992b. A canonical decomposition theory for metrics on a finite set. *Adv. Math.* 92, 47–65.

Barthélémy, J.P., and Guénoche, A. 1991. *Trees and proximity representations*, Wiley, New York.

Bryant, D., and Moulton, V. 2002. NeighborNet: An agglomerative method for the construction of planar phylogenetic networks, *in* R. Guigo, D. Gusfield, eds., 2^{nd} *Workshop on Algorithms in Bioinformatics*, 375–391, LNCS 2452, Springer.

Bryant, D., and Waddell, P. 1998. Rapid evaluation of least-squares and minimum-evolution criteria on phylogenetic trees. *Mol. Biol. Evol.* 15, 1346–1359.

Buneman, P. 1974. A note on metric properties of trees. *J. Comb. Theory B.* 17, 48–50.

Doolittle, W.F. 1999. Phylogenetic classification and the universal tree. *Science* 284, 2124–2128.

Felsenstein, J. 1997. An alternating least-squares approach to inferring phylogenies from pairwise distances. *Syst. Zool.* 46, 101–111.

Gascuel, O. 1997a. BIONJ: An improved version of the NJ algorithm based on a simple model of sequence data. *Mol. Biol. Evol.* 14, 685–695.

Gascuel, O. 1997b. Concerning the NJ algorithm and its unweighted version, UNJ, *in* B. Mirkin, F. R. McMorris, F. Roberts, and A. Rzhetsky, eds., *Mathematical Hierarchies and Biology, DIMACS Series in Discrete Mathematics and Theoretical Computer Science*, 149–170, Providence, RI.

Gascuel, O. 2000. Data model and classification by trees: The minimum variance reduction (MVR) method. *J. Classification* 17, 67–99.

Guindon, S., and Gascuel, O. 2002. Efficient biased estimation of evolutionary distances when substitution rates vary across sites. *Mol. Biol. Evol.* 19, 534–543.

Hallet, M.T., and Lagergren, J. 2001. Efficient algorithms for lateral gene transfer problems. Proc. 5th Ann. Int. Conf. on Computational Molecular Biology.

Hein, J. 1993. A heuristic method to reconstruct the history of sequences subject to recombination. *J. Mol. Evol.* 36, 396–405.

Huson, D.H. 1998. SplitsTree: A program for analyzing and visualizing evolutionary data. *Bioinformatics* 141, 68–73.

Jukes, T.H., and Cantor, C.R.. 1969. Evolution of protein molecules, *in* H.N. Munro, ed., *Mammalian Protein Metabolism*, 21–132, Academic Press, New York.

Kimura, M. 1981. Estimation of evolutionary distances between homologous nucleotide sequences. *Proc. Natl. Acad. Sci. USA* 78, 454–458.

Kuhner, M.K., and Felsenstein, J. 1994. A simulation comparison of phylogeny algorithms under equal and unequal evolutionary rates. *Mol. Biol. Evol.* 11, 459–468.

Legendre, P. 2000. Biological applications of reticulation analysis. *J. Classification* 17, 153–157.

Legendre, P., and Makarenkov, V. 2002. Reconstruction of biogeographic and evolutionary networks using reticulograms. *Syst. Biol.* 51, 199–216.

Li, W.H. 1998. *Molecular Evolution*, Sinauer, Boston.

Lockhart, P.J., Penny, D., Hendy, M.D., and Lakrum, A.W.D. 1993. Is *Prochlorothrix hollandica* the best choice as a prokaryotic model for higher plant chl-a/b photosynthesis? *Photosynthesis Res.* 73, 61–68.

Lockhart, P.J., Steel, M.A., Hendy, M.D., and Penny, D. 1994. Recovering evolutionary trees under a more realistic model of sequence evolution. *Mol. Biol. Evol.* 11, 605–612.

Makarenkov, V. 2001. T-Rex: Reconstructing and visualizing phylogenetic trees and reticulation networks. *Bioinformatics* 17, 664–668.

Makarenkov, V., and Leclerc, B. 1997. Tree metrics and their circular orders: Some uses for the reconstruction and fitting of phylogenetic trees, *in* B. Mirkin, F.R. McMorris, F. Roberts, and A. Rzhetsky, eds., *Mathematical Hierarchies and Biology, DIMACS Series in Discrete Mathematics and Theoretical Computer Science*, 183–208, American Mathematical Society, Providence, RI.

Makarenkov, V., and Leclerc, B. 1999. An algorithm for the fitting of an additive distance according to a weighted least-squares criterion. *J. Classification* 16, 3–27.

Makarenkov, V., and Leclerc, B. 2000. Comparison of additive trees using circular orders. *J. Comp. Biol.* 7, 731–744.

Margulis, L. 1981. *Symbiosis in Cell Evolution*, Freeman, San Francisco, CA.

McDade, L. 1995. Hybridization and phylogenetics, *in* P.C. Hoch and A.G. Stephenson, eds., *Experimental and Molecular Approaches to Plant Biosystematics*, 305–331, Monographs in Systematic Botany from the Missouri Botanical Garden.

Nakhleh, L., Sun, J., Warnow, T., Linder, R., Moret, B.M.E., and Tholse, A. 2003. Towards the development of computational tools for evaluating phylogenetic network reconstruction methods. *Proc. 8th Pac. Symp. on Biocomputing*, 315–326.

Rieseberg, L.H., and Ellstrand, N.C. 1993. What can molecular and morphological markers tell us about plant hybridization? *Crit. Rev. Plant Sci.* 12, 213–241.

Rieseberg, L.H., and Morefield, J.D. 1995. Character expression, phylogenetic reconstruction, and the detection of reticulate evolution, *in* P.C. Hoch and A.G. Stephenson, eds., *Experimental and Molecular Approaches to Plant Biosystematics*, 333–353, Monographs in Systematic Botany from the Missouri Botanical Garden.

Rissanen, J. 1978. Modeling by shortest data description, *Automatica* 14, 465–471.

Robinson, D.R., and Foulds, L.R. 1981. Comparison of phylogenetic trees. *Math. Biosci.* 53, 131–147.

Posada, D., and Crandall, K.A. 2001. Intraspecific phylogenetics: Trees grafting into networks. *Trends Ecol. Evol.* 16, 37–45.

Saitou, N., and Nei, M. 1987. The neighbor-joining method: A new method for reconstructing phylogenetic trees. *Mol. Biol. Evol.* 4, 406–425.

Sattath, S., and Tversky, A. 1977. Phylogenetic similarity trees. *Psychometrika* 42, 319–345.

Sawyer, S. 1989. Statistical tests for detecting gene conversion. *Mol. Biol. Evol.* 6, 526–536.

Smouse, P.E. 2000. Reticulation inside species boundary. *J. Classification* 17, 165–173.

Sneath, P.H.A. 2000. Reticulate evolution in bacteria and other organisms: How can we study it? *J. Classification* 17, 159–163.

Sneath, P.H.A., Sackin, M.J., and Ambler, R.P. 1975. Detecting evolutionary incompatibilities from protein sequences. *Syst. Zool.* 24, 311–332.

Steel, M.A. 1994. Recovering a tree from the leaf colorations it generates under a Markov model. *Appl. Math. Lett.* 72, 19–24.

Stephens, J.C. 1985. Statistical methods of DNA sequence analysis: Detection of intragenic recombination or gene conversion. *Mol. Biol. Evol.* 2, 539–556.

Stevenson, R.J., Bothwell, M.L., and Lowe, R.L. 1996. *Algal Ecology*, Academic Press, San Diego.

Swofford, D.L., and Olsen, G.L. 1996. Phylogeny reconstruction, *in* D.M. Hill, ed., *Molecular Systematics*, 407–514, Sinauer.

Address correspondence to:
Vladimir Makarenkov
Département d'informatique
Université du Québec à Montréal
C.P. 8888
Succ. Centre-Ville
Montréal (Québec), Canada

E-mail: makarenkov.vladimir@uqam.ca