Syst. Biol. 51(2):199–216, 2002

Reconstruction of Biogeographic and Evolutionary Networks Using Reticulograms

PIERRE LEGENDRE[1] AND VLADIMIR MAKARENKOV[1,2]

[1]*Département de sciences biologiques, Université de Montréal, C.P. 6128, succursale Centre-ville, Montréal, Québec, Canada H3C 3J7; E-mail: Pierre.Legendre@umontreal.ca*
[2]*Institute of Control Sciences, 65 Profsoyuznaya, Moscow 117806, Russia; E-mail: Makarenkov.Vladimir@uqam.ca*

Abstract.—A reticulogram is a general network capable of representing a reticulate evolutionary structure. It is particularly useful for portraying relationships among organisms that may be related in a nonunique way to their common ancestor—relationships that cannot be represented by a dendrogram or a phylogenetic tree. We propose a new method for constructing reticulograms that represent a given distance matrix. Reticulate evolution applies first to phylogenetic problems; it has been found in nature, for example, in the within-species microevolution of eukaryotes and in lateral gene transfer in bacteria. In this paper, we propose a new method for reconstructing reticulation networks and we develop applications of the reticulate evolution model to ecological biogeographic, population microevolutionary, and hybridization problems. The first example considers a spatially constrained reticulogram representing the postglacial dispersal of freshwater fishes in the Québec peninsula; the reticulogram provides a better model of postglacial dispersal than does a tree model. The second example depicts the morphological similarities among local populations of muskrats in a river valley in Belgium; adding supplementary branches to a tree depicting the river network leads to a better representation of the morphological distances among local populations of muskrats than does a tree structure. A third example involves hybrids between plants of the genus *Aphelandra*. [Biogeographic history; distance matrix; phylogenetic tree; reticulation network; reticulogram.]

Reticulate evolution refers to evolutionary processes that cannot be fully represented by the tree model. Reticulate patterns of relationships are found in nature in the following phylogenetic situations: (1) lateral gene transfer in bacterial evolution, which can be studied either in the deep phylogeny or in presently evolving groups; (2) hybridization between species, including allopolyploidy in plants; (3) microevolution of local populations within a species, involving genetic differentiation of allopatric populations, gene exchange through migration, or both (example 2 below); (4) homoplasy, the portion of phylogenetic similarity resulting from evolutionary convergence (i.e., parallel evolution and reversals), which can be represented by reticulation branches added to a phylogenetic tree; and nonphylogenetic situations, such as (5) host–parasite relationships involving host transfer and (6) vicariance and dispersal biogeography (example 1 below).

A reticulogram is a type of graph capable of representing relationships among organisms that may have more than one path connecting an organism to another. Such a structure, which contains cycles, cannot be represented by a phylogenetic tree, which is acyclic by definition. Phylogenetic trees are particular cases of reticulograms and include the extra property that the path from the root to any object is unique.

Sneath (1975) summarized the biological evidence from various fields and first showed how reticulate evolution could be represented using modified cladograms. The biological concepts that form the foundation of reticulation analysis as well as the methods currently available for the reconstruction of reticulograms have been summarized in a special section of the *Journal of Classification* (Legendre, 2000a).

Here we propose a new method for reconstructing reticulation networks and apply the reticulate evolution model to ecological biogeographic, population microevolutionary, and hybridization problems. Other examples involving reticulation events that can be interpreted in terms of hybridization, homoplasy, and endosymbiosis are developed in a companion paper (Makarenkov and Legendre, in prep.).

INFERRING PHYLOGENETIC TREES AND RETICULATION NETWORKS

A matrix of pairwise distances among leaves can be associated with any phylogenetic tree (see Zaretskii, 1965; Buneman, 1974;

Lapointe and Legendre, 1992). Such a matrix is called a tree distance matrix. Let d_{ij} denote the value of the tree distance between a pair of taxa i and j corresponding to leaves i and j of the tree. The condition known as the four-point condition or additive inequality, which characterizes phylogenetic trees, is the following:

$$d_{ij} + d_{kl} = \max\{d_{ik} + d_{jl};\ d_{il} + d_{jk}\}, \quad \text{for (all } i, j, k, \text{ and } l) \qquad (1)$$

A phylogenetic tree uniquely defines a set of distances d that satisfy the four-point condition. Inversely, whenever a set of distances d satisfies this condition, it defines a unique phylogenetic tree.

Because in practice raw empirical data rarely satisfy the four-point condition, a phylogenetic tree representation has to be inferred by using an appropriate fitting method. A wide variety of methods allow reconstruction of phylogenetic trees from a given distance matrix; for an overview, see, for example, Barthélemy and Guénoche (1991) or Swofford et al. (1996). Each method approximates the observed distance matrix according to a stated criterion. The most popular estimation criteria are (weighted) least squares, neighbor-joining, maximum likelihood, and maximum parsimony.

We will use the least-squares criterion in the present study: it is appropriate for the task; it is the most widely used criterion for statistical estimation; and it is often computationally faster than methods based on parsimony or maximum likelihood. The problem of fitting a phylogenetic tree to a distance matrix according to least squares was shown to be NP-hard by Day (1987, 1996). This result has stimulated the development of several heuristic approaches allowing inference of tree topologies in polynomial time.

Several authors have proposed algorithms for the representation of empirical distances among taxa by using general network models instead of phylogenetic trees (for example, Feger and Bien, 1982; Feger and Droge, 1984; Orth, 1988; Klauer, 1988, 1989; Klauer and Carroll, 1989, 1995). In such network models, the taxa are represented by the nodes of a weighted graph. The minimum path-length distances between pairs of taxa approximate the empirical distances. For example, Klauer and Carroll (1989, 1995) designed an algorithm for constructing a least-squares representation of a distance matrix by a general network with a fixed number of branches, a constraint that would not be appropriate for reticulation analysis. Their procedure fits a network with a specified number of branches so that the minimum path-length distances optimally approximate the observed data. Readers are referred to De Soete and Carroll (1996) for an overview of the general network fitting techniques.

Lapointe (2000) reviewed four distance-based methods that can be used to account for reticulation events. Pyramids (Diday and Bertrand, 1986) and weak hierarchies (Bandelt and Dress, 1989) are techniques developed to fit dendrograms with overlapping clusters; they can be used for classification but are ill-adapted to phylogenetic analysis. The concept of weak clusters, leading to the construction of a weak hierarchy for an empirical similarity matrix, was proposed by Bandelt and Dress (1989). Their weak hierarchy clustering technique can be used to represent an additive tree structure that contains some ambiguous solutions. Weak hierarchies allow one to refrain from resolving conflicting features right away, as would be required when producing a standard dendrogram. Subsequent investigation of weak clusters by Bandelt and Dress (1992a, 1992b), Bandelt (1995), and Dress et al. (1996) has given rise to the popular method of split decomposition. In this transformation-based approach, the observed data are canonically decomposed into a sum of "weakly compatible splits" and represented by a so-called splitsgraph. For perfect phylogenetic data, the splitsgraph is a phylogenetic tree; less perfect data are depicted with a tree-like network that represents the conflicting information contained in the data. In a splitsgraph, a pair of nodes may be linked by a set of parallel branches representing alternative solutions. Splitsgraphs are mostly used to display incompatibilities in data sets. However, this is often not the purpose of reticulate analysis, except when the study focuses on displaying homoplasy. An example comparing splitsgraph analysis with reticulogram analysis, as introduced in the present paper, is presented by Makarenkov and Legendre (in prep.).

Here we describe a new way of modeling the representation of an empirical distance matrix. A reticulogram is based on a network topology which includes a set of nodes labeled with taxon, locality, or some other form of object names, as well as a set of intermediate nodes. Reticulograms may, for instance, describe the fine-scale spatial structure of ecological populations or the broad-scale spatial structure of species assemblages studied in historical biogeography, which results from migrations that occurred across historical time scales. The term reticulogram, created by Lefkovitch in 1983 (cited in Legendre, 1984), is a condensation of, and stands for, reticulated cladogram (Wanntorp, 1983). The distance between a pair of nodes in a reticulogram is defined as the minimum path-length distance over the set of all paths linking them. In Figure 1 (left) for instance, the distance between objects x and z is 3, whereas in Figure 1 (right) the minimum path-length distance, which goes through the branch between x and z, is 0.5. The least-squares fitting of an optimal reticulogram structure representing a given distance measure has been shown to be a delicate problem; this problem is at least as complicated as the NP-hard problem of fitting a phylogenetic tree to a given distance matrix by least squares.

We decided to use a heuristic algorithm, which runs in polynomial time relative to the number of observed objects (taxa and so forth), to seek an optimal reticulogram representation of a given distance matrix. Given that the problem of inferring a phylogenetic tree from evolutionary distances is a very well studied issue, and taking into account the fact that several efficient tree-fitting algorithms are already available, we propose to start the reticulogram reconstruction procedure from a phylogenetic tree topology that provides an initial fit for the distance matrix. In this approach we add new branches, called reticulation branches, one at a time to a growing network structure, minimizing at each step the least-squares loss function, which is computed as the sum of the quadratic differences between the values of the distance matrix and the associated reticulogram estimates.

The method of continuous track analysis (CTA), proposed by Alroy (1995) to depict reticulate patterns in phylogenetic and biogeographic studies, bears some resemblance to our method. CTA is a parsimony method used primarily with paleontological material. The inner nodes of the tree represent fossils in the data set subjected to the analysis, so that no internal nodes of unknown identity have to be added to the network. As a consequence, just as in reticulogram analysis discussed in this paper, an initial tree is computed (in CTA, the Wagner method is used to obtain the starting tree); then some connecting branches are removed and replaced by connections that minimize the number of "track fragments," which represent continuity of character states through the phylogeny. CTA cannot be used in phylogenetic problems in which the inner nodes are unknown taxa.

Several methods have been proposed to detect reticulate evolutionary events in nucleotide sequence data (e.g., Jakobsen and Easteal, 1996). These methods also have their limitations; for example, they cannot be used to analyze biogeographic problems.

In particular cases, reticulation branches may be chosen by a supplementary matrix of weights or using any supplementary constraint matrix associated with a given distance matrix. For instance (example 1 below), we constructed a spatially constrained reticulogram representing the postglacial dispersal of freshwater fishes in the Québec peninsula.

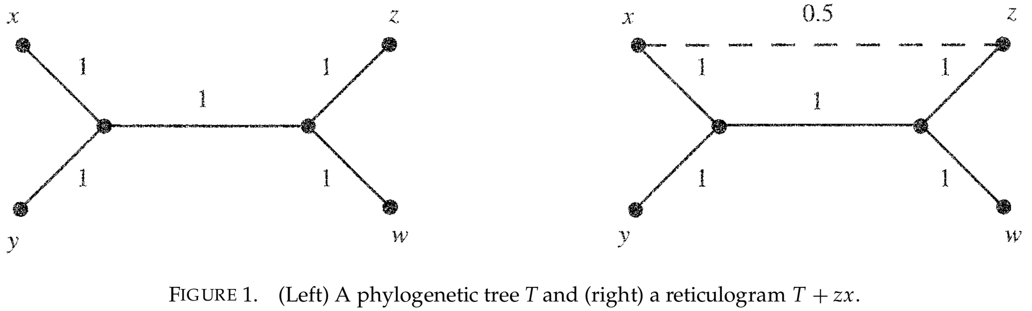

FIGURE 1. (Left) A phylogenetic tree T and (right) a reticulogram $T + zx$.

A New Method for Reticulogram Reconstruction

Basic Definitions

Let us introduce some basic definitions. A reticulogram or reticulation network R is a triplet (N, B, l) such that N is a set of nodes, B is a set of branches, and l is a function of branch lengths that assigns real nonnegative numbers to the branches. Each node i is either a taxon belonging to a set X or an intermediate node belonging to $N - X$.

A reticulogram is connected if, for every pair of nodes i and j, there is at least one path from i to j. The reticulogram is called undirected if no direction is associated with the branches. Given a connected and undirected reticulogram R, the minimum path-length distance between nodes i and j, denoted r_{ij}, is defined as follows:

$$r_{ij} = \min\{l_p(i,j) \mid p \text{ is a path from } i \text{ to } j\} \qquad (2)$$

A set of reticulation distances, denoted r, 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 reticulogram.

Selected Approach

In this section we describe an algorithm for inferring a connected and undirected reticulogram from an empirical distance matrix. First, we reconstruct a phylogenetic tree from a distance matrix, using one of the existing fitting algorithms, for example, neighbor-joining (Saitou and Nei, 1987), Fitch (Felsenstein, 1997), or weighted least squares (Makarenkov and Leclerc, 1999). Supplementary branches are then added to the phylogenetic tree, one at a time, each one minimizing a least-squares or weighted least-squares loss function. Addition of reticulation branches stops when the minimum of a stopping criterion is reached. The stopping criteria we are using (see Eqs. 5 and 6) take into account the least-squares loss function as well as the number of parameters of the reticulogram under construction. Because in our study the reticulogram technique is based on the least-squares loss function, we have used an initial phylogenetic tree for which topology and branch lengths are also determined by least squares rather than by the parsimony or maximum likelihood criteria. Methods of reticulation analysis could be developed for parsimony (see Alroy, 1995) or maximum likelihood, but they would involve entirely different algorithms.

Let $\mathbf{D}$ be a distance matrix on the finite set X of n taxa and let T be a phylogenetic tree inferred from $\mathbf{D}$ by means of an available tree-fitting method. This tree has at most n leaves and $2n - 3$ branches. Such a nondegenerate tree is called a binary tree. Any phylogenetic nonbinary tree can be transformed into a binary tree associated with the same tree distance matrix, by adding to the nonbinary tree branches of null length where appropriate. In this study, we consider binary phylogenetic trees as the foundation for the reticulogram reconstruction algorithm. Thus, reticulation networks introduced in this paper will always comprise $2n - 2$ nodes, including $n - 2$ intermediate nodes and n leaves labeled according to the taxa in X. The number of branches in a reticulogram will vary from $2n - 3$, which is the number of branches in a phylogenetic tree, to $(2n - 2)(2n - 3)/2$, which is the number of branches in a complete graph with $2n - 2$ nodes. The original tree may be rooted or not; this does not really matter when constructing undirected reticulograms.

Mathematical Description of the Problem

Let us now discuss the problem from a mathematical point of view. Our task is to reconstruct a connected and undirected reticulation network R with a fixed number (say, K) of links that represents best, according to least squares, a given distance matrix $\mathbf{D}$. The optimization problem at stake can be formulated as follows:

$$Q = \sum_{i \in X} \sum_{j \in X} (d_{ij} - r_{ij})^2 \rightarrow \min \qquad (3)$$

under the following constraints: $r_{ij} = 0$, for all $i,j \in X$, and reticulation distance r is associated with a reticulogram R with K branches. In many instances, a phylogenetic tree already represents a good estimate of an evolutionary structure; several efficient tree-fitting algorithms are available in the literature. We decided to start reticulogram reconstruction from a phylogenetic tree topology inferred from a given distance matrix $\mathbf{D}$. This was also the strategy followed by

Alroy (1995). New reticulated branches will be added to the growing reticulogram, one during each iteration of the algorithm.

Consider a binary phylogenetic tree T inferred from a distance matrix $\mathbf{D}$ and a pair of nodes x and y in T that are not linked by a branch. In the first iteration, we have to add a first reticulation branch to T. Thus, we have to find an optimal value l, according to the least-squares objective function Q (Eq. 3), for a new potential branch xy that we are adding to the tree T, while keeping fixed the lengths of all existing branches (Fig. 2). We will not present here the rather lengthy and not very illuminating algebra needed for identifying the particular value that optimizes the length of the new branch. In summary, we consider several length values, using a quadratic spline as the criterion function. We optimize within the various intervals between spline points to obtain the optimal length value l for a particular branch xy.

When this optimum value has been found, we compute the corresponding value of the objective function, Q_{xy}, which provides the gain in fit after branch xy has been added, and retain the value Q_{xy} as a potential minimum of the least-squares criterion after addition of the first reticulation branch. Then we select another pair of objects in T (say, i and j) that are not linked by a branch and compute both the optimal value of a potential new branch ij and the corresponding gain in fit Q_{ij}. If Q_{ij} is less than Q_{xy}, we retain ij as the new potential branch to be added to T during the first iteration.

To obtain the optimum value of the least-squares criterion Q over the set of all possible new branches, this computation should be repeated for all pairs of tree nodes that are not linked by a branch. The exact solution can be found in polynomial time. One will need $O(n^4)$ time to optimally place a new branch into a phylogenetic tree with n leaves. Similarly, in the second and forthcoming iterations, an extra reticulation branch, the one contributing the most to the reduction of the least-squares function Q, will be added to the reticulation network. In a later subsection, we describe goodness-of-fit criteria that can be used to stop the process of adding new branches to the reticulogram.

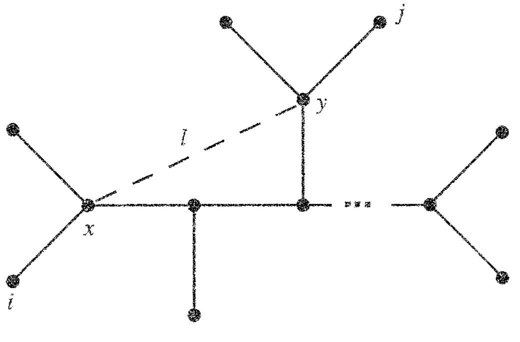

FIGURE 2. A new branch of length l can be added to the tree T between nodes x and y.

Weighted Least-Squares Criterion

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

$$Q = \sum_{i \in X} \sum_{j \in X} w_{ij}(d_{ij} - r_{ij})^2 \rightarrow \min \quad (4)$$

The new feature here is the weight function w_{ij} applied to the separation of taxa i and j. The function w is symmetric, taking nonnegative values.

The weighted least-squares criterion may be useful in various practical situations. If some values of an observed distance measure are known to be uncertain, this information can be incorporated in the weighted least-squares loss function by giving low values to the weights corresponding to the uncertain entries. If some values in the distance matrix are unknown, these missing data can be handled by setting the associated weights to 0. In problems involving spatially autocorrelated data, such as biogeography, weights may be given as an inverse function of the variogram, a spatial autocorrelation function originally developed in the field of geostatistics; this way, larger distances receive lower weights.

In the present study, we used the weighted form of the algorithm to impose a constraint of spatial contiguity to the freshwater fish data from the Québec peninsula (example 1 below). Weights of 0 were given to pairs of regions that did not have common boundaries, whereas weights of 1 were assigned to pairs of adjacent regions. In that study, we did not use weights other than 0 and 1. In a biogeographic studies in general, a noninteger weight w_{ij} may represent a probability for regions i and j to be linked, for example, by a river connection.

For an overview of applications of the weighted least-squares criterion in phylogenetics, see Swofford et al. (1996). Numerous effective methods exist for inferring phylogenetic trees by using weighted least squares. Felsenstein (1997) described how this kind of optimization is performed in the package PHYLIP; see also the recent papers by Makarenkov and Leclerc (1999) and Gascuel (2000). Bryant and Waddell (1998) and Makarenkov and Leclerc (1999) describe how to compute optimal branch lengths for a tree with fixed topology in the weighted case.

The procedure for computing the optimal length l of a new hypothetical branch xy can be adapted to the weighted least-squares objective function. As in the unweighted case, the optimal solution can be obtained in polynomial time relative to the number of taxa.

Some Properties of a Reticulation Distance

A reticulation distance is no longer a tree distance. Figure 1 gives an example of a reticulation distance not satisfying the four-point condition characterizing phylogenetic trees. The four-point condition is fulfilled if the greatest two among the three possible sums of distance are equal. This condition holds for the tree distanc δ associated with tree T (Fig. 1 left) but not for the distance r associated with the reticulogram $T + xy$ (Fig. 1 right):

$$\delta(x,y) + \delta(z,w) = 4; \quad r(x,y) + r(z,w) = 4;$$

$$\delta(x,z) + \delta(y,w) = 6; \quad r(x,z) + r(y,w) = 3.5;$$

$$\delta(x,w) + \delta(y,z) = 6; \quad r(x,w) + r(y,z) = 5.$$

However, because the reticulation distance is the minimum path-length distance in a weighted graph, it satisfies the triangle inequality.

Stopping Rules for Adding Reticulation Branches

A reticulogram contains more branches and consequently uses more parameters than a (bi)furcating phylogenetic tree. As in all statistical models, including more parameters means a better fit but a loss in simplicity and robustness; we deal with this issue by using a special cost criterion. In this subsection, we consider two goodness-of-fit criteria to measure the improvement in fit when reticulation branches are added to a tree. These criteria take into account the least-squares loss function, which can be weighted or not, as well as the number of parameters of the reticulogram. The minima of these criteria provide stopping rules for adding reticulation branches when the exact number of branches is unknown in advance, as is often the case in biological applications.

The maximum number of branches one might place into a reticulogram obtained from a basic phylogenetic tree with n leaves is $(2n-2)(2n-3)/2$. However, we know that any metric distance can be represented by a complete graph with $n(n-1)/2$ branches. We can thus consider this last value the maximum possible number of branches in a reticulogram. Consequently, the number of degrees of freedom of a reticulogram with N branches may be defined as $[n(n-1)/2] - N$; accordingly, the first function we propose to consider is the following:

$$Q_1 = \frac{\sqrt{\sum_{i\in X}\sum_{j\in X}(r_{ij} - d_{ij})^2}}{[n(n-1)/2] - N} = \frac{\sqrt{Q}}{[n(n-1)/2] - N}. \tag{5}$$

An interesting feature of this criterion is that in our algorithm 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 may be used for stopping the addition of reticulation branches to the growing reticulogram.

The least-squares loss function itself, instead of its square root, may be considered as an appropriate numerator for the goodness-of-fit criterion. Q_2 is a slightly modified criterion whose minimum is usually achieved several iterations later than the minimum of the function Q_1.

$$Q_2 = \frac{\sum_{i\in X}\sum_{j\in X}(r_{ij} - d_{ij})^2}{(n(n-1)/2 - N)} = \frac{Q}{(n(n-1)/2 - N)}. \tag{6}$$

As a consequence, the modified criterion usually requires adding more reticulation branches to the network than does criterion Q_1.

During a Monte Carlo study described below, we carried out an investigation to explore how many local minima the criteria

Q_1 and Q_2 may possess over the interval $[2n - 3, (2n - 3) + 2N^*]$ of possible values of N, where N^* is the number of reticulation branches yielding the first local minimum of Q_1 or Q_2. The upper bound of $(2n - 3) + 2N^*$ chosen for these simulations was likely to contain the number of branches providing the global minimum of Q_1 or Q_2. In addition to the global minimum, criterion Q_1 had local minima only in 0.972% of cases (i.e., for 35 out of 3,600 basic phylogenetic trees on the sets of 10, 20, and 30 taxa analyzed; see the simulation study for more details), whereas criterion Q_2 had no local minima at all over the observed interval of values of N. The absence of local minima for Q_2 and the very small percentage of appearance of local minima for Q_1 justifies the usage of the first minimum of these functions as a stopping criterion for addition of reticulation branches. In the section of practical examples, we will illustrate the application of stopping rules Q_1 and Q_2 in real-world situations.

Another interesting stopping rule, which will be used in the Monte Carlo simulations (below), consists in adding to the phylogenetic tree several reticulations that provide a given percentage of gain in fit, compared with the least-squares loss Q of the unreticulated tree. In the simulations, we explored how many reticulation branches were added by the algorithm to produce a 10%, 15%, or 25% gain in fit with respect to the unreticulated tree.

Algorithmic Time Complexity

For an $(n \times n)$ distance matrix, the algorithm described above requires at most $O(kn^4)$ time to add k reticulation branches to the phylogenetic tree. The algorithm requires at most $O(n^2)$ time to compute the optimal length of a new hypothetical branch xy. Considering that the possible number of pairs of nodes x and y not linked by a branch in a reticulogram under construction is of the order of n^2, we conclude that the algorithm requires $O(kn^4)$ time to place k branches onto the reticulogram. However, in our simulation study, the algorithm required on average $O(n)$ time to compute the optimal length of a new branch xy. Consequently, in practice, the overall time complexity of placing k branches to the reticulogram is closer to $O(kn^3)$ than to $O(kn^4)$.

Simulation Study

In carrying out simulations for the algorithm described in the previous section, we used the evaluation approach proposed by Pruzansky et al. (1982) to compare phylogenetic tree-fitting algorithms, adapting their strategy to test reticulogram reconstruction algorithms.

Each simulated data set was obtained as follows: First, an unrooted tree topology with n leaves and $2n - 3$ branches was randomly generated. For each such tree topology, the length of each branch was selected at random from a uniform distribution on the real interval [0,1], leading to a phylogenetic tree T. A tree distance t, corresponding to the obtained tree T, was computed. To simulate a reticulogram, we then placed in T a random number (sampled from a uniform distribution on the interval $[1,n]$ of integers) of branches, with random lengths selected as above. The locations of these branches were also selected randomly. The reticulation distance matrix $\mathbf{R}$ associated with the constructed reticulogram R, computed by using minimum path-length distances between pairs of taxa, was normalized to have unit variance.

Normally distributed random errors with mean 0 and variances $\sigma^2 = \{0.0, 0.1, 0.25, 0.5\}$ were added to $\mathbf{R}$ to obtain replicates of the distance matrix $\mathbf{D}$. In the rare cases where a negative value of $d(x, y)$ arose, it was replaced with the constant 0.01. We carried out simulations with matrices of size (10×10), (20×20), and (30×30). Thus, we effectively created "phylogenetic tree + reticulation branches + noise" realizations. Because our procedure does not guarantee reconstruction of a true reticulogram when the data are indeed reticulation distances, we also carried out similar tests when the random noise variance σ^2 was 0. For each combination of values (n, σ^2), 100 data sets were generated. The results reported in Tables 1 and 2 are averages from 100 different simulated distance matrices, for a total of 3,600 simulated data sets in each table.

Because our method includes a method for fitting phylogenetic trees as its first part, we carried out a Monte Carlo investigation involving three different tree reconstruction methods that may be used to compute the basic phylogenetic tree structure and the associated tree distance matrix $\mathbf{\Delta}$. The following phylogenetic tree fitting methods were used:

TABLE 1. Average proportions of variance of the distance matrices accounted for by the tree reconstruction algorithms (columns *V.Tree%*) and the reticulogram reconstruction procedure (columns *V.Ret%*) for different amounts of variance (σ^2) of the random error. Larger values of V are better.

	Tree	$\sigma^2 = 0.0$		$\sigma^2 = 0.1$		$\sigma^2 = 0.25$		$\sigma^2 = 0.5$	
n	algorithm	*V.Tree%*	*V.Ret%*	*V.Tree%*	*V.Ret%*	*V.Tree%*	*V.Ret%*	*V.Tree%*	*V.Ret%*
10	ADDTREE	94.49	97.18	87.82	91.95	84.17	88.22	75.88	80.98
10	NJ	94.58	97.20	88.06	92.09	84.27	88.26	76.09	81.03
10	MW	94.69	97.25	88.21	92.15	84.51	88.36	76.42	81.29
20	ADDTREE	89.56	94.61	83.12	88.55	76.74	81.85	66.77	72.40
20	NJ	89.44	94.52	83.36	88.54	76.82	81.86	66.61	72.28
20	MW	90.56	94.94	84.30	88.90	77.56	82.29	67.40	72.95
30	ADDTREE	87.28	93.47	81.72	87.07	72.29	78.48	63.17	69.26
30	NJ	87.44	93.49	81.69	87.02	72.10	78.31	63.00	69.15
30	MW	88.98	94.13	82.70	87.43	73.44	78.91	64.16	69.95

NJ, neighbor-joining; MW, method of weighted least squares.

the ADDTREE method of Sattath and Tversky (1977);
the neighbor-joining method of Saitou and Nei (1987); and
the weighted least-squares method of Makarenkov and Leclerc (1999).

Goodness-of-fit was estimated by the following two quantities, computed for all simulated data sets:

1. The proportion of variance accounted for by the phylogenetic trees or the reticulograms, as expressed in Eq. 7, where *m(d)* is the mean value in the upper-triangular portion of the distance matrix **D**, and r is the fitted reticulation distance (or the tree distance):

$$Var\% = 100 \times \left(1 - \frac{\sum_{ij \in X}(d_{ij} - r_{ij})^2}{\sum_{ij \in X}(d_{ij} - m(d))^2}\right) \tag{7}$$

   This quantity was computed for the tree distance matrix obtained by using one of the three above-mentioned tree-fitting methods (columns *V.Tree%* in Table 1) as well as for the reticulation distance provided by our algorithm (columns *V.Ret%* in Table 1). The procedure of adding reticulation branches was stopped when k new branches had been placed onto the tree network, where k is the random number of reticulation branches in the true reticulogram generated before adding noise.
2. The criterion of goodness-of-fit Q_1 (Eq. 5), which takes into account the least-squares loss as well as the number of degrees of freedom of a reticulogram (or tree) with N branches. The quantities Q_1.*Ret* reported in Table 2 represent the averages of the minimum values of this function achieved on the interval of integer values $[1 \ldots k]$, where k is a random number of reticulation branches placed in the true reticulogram before adding the noise. The same quantities were also computed for all the basic phylogenetic trees and reported in the columns Q_1.*Tree* of Table 2.

TABLE 2. Mean values of the goodness-of-fit criterion Q_1 computed for the tree reconstruction algorithms (columns Q_1.*Tree*) and the reticulogram reconstruction algorithm (columns Q_1.*Ret*) for different amounts of variance (σ^2) of the random error. Smaller values of Q_1 are better.

	Tree	$\sigma^2 = 0.0$		$\sigma^2 = 0.1$		$\sigma^2 = 0.25$		$\sigma^2 = 0.5$	
n	algorithm	Q_1.*Tree*	Q_1.*Ret*	Q_1.*Tree*	Q_1.*Ret*	Q_1.*Tree*	Q_1.*Ret*	Q_1.*Tree*	Q_1.*Ret*
10	ADDTREE	0.0484	0.0400	0.0845	0.0792	0.1034	0.1001	0.1384	0.1357
10	NJ	0.0480	0.0399	0.0837	0.0787	0.1031	0.0998	0.1379	0.1354
10	MW	0.0476	0.0397	0.0832	0.0782	0.1024	0.0993	0.1369	0.1342
20	ADDTREE	0.0275	0.0209	0.0380	0.0336	0.0479	0.0451	0.0623	0.0600
20	NJ	0.0275	0.0211	0.0378	0.0336	0.0478	0.0451	0.0625	0.0601
20	MW	0.0262	0.0203	0.0368	0.0331	0.0471	0.0446	0.0617	0.0594
30	ADDTREE	0.0186	0.0137	0.0245	0.0215	0.0322	0.0295	0.0405	0.0386
30	NJ	0.0184	0.0137	0.0245	0.0216	0.0323	0.0296	0.0406	0.0387
30	MW	0.0172	0.0130	0.0238	0.0213	0.0314	0.0292	0.0400	0.0382

NJ, neighbor-joining; MW, method of weighted least squares.

The Monte Carlo simulations lead to the following observations:

1. The better the fit provided by a tree reconstruction method, the closer the reticulation distance was to the generated distance. This observation remains true for the fit measured either by the percentage of variance accounted for (Table 1) or by the goodness-of-fit criterion Q_1 (Table 2). In these simulations, the weighted least-squares tree reconstruction method generally provided a better fit than the neighbor-joining and ADDTREE methods, which gave results very similar to each other.
2. For any considered combination of parameters (n, σ^2), reticulograms always provided better average values of the goodness-of-fit criterion Q_1 than did phylogenetic trees, regardless of the basic tree reconstruction method (Table 2).
3. The smaller the noise or the size of the distance matrix, the better the percentage of variance accounted for (Table 1) and the criterion of goodness-of-fit Q_1 (Table 2) for the tree reconstruction methods and the reticulograms.

Further simulations were carried out to investigate the behavior of the algorithm for trees that did not contain reticulation branches. No reticulation branch should be added when the data consist of unreticulated additive trees without error. In the presence of error, however, we can expect some reticulation branches to be formed, as in the type I error of statistical tests.

Random trees, with $n = \{10, 20, 30\}$, were generated as described above, and the corresponding patristic distance matrices were computed. Normally distributed random errors were with mean 0 and variances $\sigma^2 = \{0.0, 0.1, 0.25, 0.5\}$. No reticulation branches were added to the trees. We fitted a tree by the method method of weighted least squares, followed by reticulation analysis. We calculated how many reticulation branches were necessary to obtain a gain in adjustment (measured by coefficient Q, Eq. 3) of 10%, 15%, and 25%, compared with the least-squares coefficient Q obtained for the fitted tree.

Simulation results for 100 random trees (Table 3) confirm that no reticulation branches were added by the algorithm when analyzing error-free data ($\sigma^2 = 0.0$). In the case of trees to which error had been added ($\sigma^2 = \{0.1, 0.25, 0.5\}$), the number of reticulation branches necessary to produce preselected increases in fit depended on the number of leaves n and the amount of error σ^2 added to the patristic distance matrix of the random tree. The number of "false reticulation branches" added to the tree by the algorithm in the case of noisy data increased with increasing n and with the amount of noise in the data. These results indicate that if reticulation analysis is used with unreticulated but noisy data, the method is likely to produce reticulation branches that represent incompatibilities because of the noisy nature of the data.

TABLE 3. Mean number of reticulation branches, in 100 simulations for type I error, needed to obtain a gain in fit (measured by the least-squares coefficient Q) of 10%, 15%, and 25%, compared with the least-squares coefficient Q of the fitted tree.

n	$\sigma^2 = 0.0$	$\sigma^2 = 0.1$	$\sigma^2 = 0.25$	$\sigma^2 = 0.5$
Gain in fit: 10%				
10	0.00	1.66	2.59	2.93
20	0.00	4.08	4.71	5.24
30	0.00	8.05	9.09	10.73
Gain in fit: 15%				
10	0.00	3.21	4.23	5.40
20	0.00	6.99	8.37	10.23
30	0.00	14.14	16.64	20.92
Gain in fit: 25%				
10	0.00	8.27	12.20	14.73
20	0.00	16.81	21.15	25.78
30	0.00	29.00	29.91	30.00[a]

[a] The maximum number of reticulation branches allowed in this simulation study was 30.

EXAMPLES

We will now explore how the reticulation method works in practice, using examples from biogeography, population microevolution, and hybridization.

Example 1: Postglacial Dispersal Routes of Freshwater Fishes

Methods designed for inferring phylogenetic trees can sometimes help reconstruct biogeographic history. This is the case when a tree-like structure of geographic dispersal is sought, describing the invasion of an area by a group of species. Most methods of phylogenetic reconstruction assume that different branches of the tree evolved independently from one another, thus preventing reticulate interactions between branches after their point of splitting-off. This assumption is unrealistic in many biogeographic problems. Considering a general network model establishing extra relationships between observed territory units would allow one to identify and represent potential reticulate interactions.

Legendre and Legendre (1984) used present-day data on fish fauna to reconstruct plausible routes taken by freshwater fishes to reinvade the Québec peninsula after the last glaciation. The Laurentide Ice Sheet retreated from the peninsula from about 14,000 years to 5,000 years ago. The presence or absence of 109 species in 289 pixels (1°-square map units) was analyzed. The territory under study was divided into 21 regions by spatially constrained clustering of the 289 pixels, based on a matrix of Sørensen similarity coefficients among pixels computed from the fish presence–absence data. These larger territory units are numbered 1 to 21 in Figure 3. The Sørensen coefficient, a similarity measure for species presence–absence data, is widely used in ecology to compare sites or regions. Spatially constrained clustering is a clustering method whereby a map containing operational geographic units (OGUs) characterized by multivariate data can be divided into regions containing contiguous OGUs (see Legendre and Legendre [1998] for details).

Using the presence–absence of the 85 stenohaline (i.e., restricted to fresh water) species only, Legendre (1986) applied several methods of phylogenetic analysis to reconstruct the postglacial dispersal routes. Among them, a Camin–Sokal phylogenetic tree (Camin and Sokal, 1965) was fitted to a matrix of the fish presence–absence in the 21 territory units (Fig. 3). When fish reinvaded the Québec peninsula at the end of the last glaciation, the root of the tree, represented by the glacial refugia (Hudson and Mississippi rivers), was hypothesized to have contained all 85 stenohaline species presently found in the peninsula—except *Moxostoma hubbsi* (Legendre, 1942), the only species of vertebrate endemic to Québec; that species appeared in the Saint Lawrence River and some affluents since the last glaciation. As the fish assemblage moved north and east, using river pathways, it could only lose species along the way; it could not recreate them. The Camin–Sokal parsimony method mimics this process in that reversals are not permitted to occur during tree reconstruction. The "ancestral state," found in the root of the tree, was thus the presence of a species; the "derived state" was taken to be the absence of a species in >50% of the pixels of a territory unit. The branches of the Camin–Sokal tree are represented by solid lines in Figure 3; internal nodes are numbered 23 to 42.

This, of course, was a simplified view. Biogeographic reticulation events most probably occurred, with fish species reaching regions through alternative river pathways. This is why the reticulogram (below) explains the reinvasion process better than the Camin–Sokal tree.

For the present analysis, the presence–absence data were transformed into a metric distance matrix by using the Jaccard coefficient. This matrix served first to find the optimal branch lengths of the Camin–Sokal tree, whose topology was taken from Legendre (1986), and second to transform the phylogenetic tree into a reticulation network. Reticulation branches (dashed lines) were added to the tree by using a spatially constrained form of our algorithm, that is, with the constraint that reticulation branches could only be drawn between adjacent territory units. Territory units that touch each other by even a single point in Figure 3 were considered adjacent.

The total sum of squares (total SS) in the Jaccard distance matrix was 5.487 before fitting the tree; after fitting the tree, the residual SS was 1.660. This result means that the 41 branches of the tree together explained 3.828 units of SS, or 0.0934 unit per branch on average. Compare this value to the contributions to SS of the nine reticulation branches added to the tree, which together explained 0.249 units of SS—from 0.0387 for the first reticulation branch (the 17–37 link) to 0.0164 for the ninth reticulation branch (the 12–13 link). By adding nine new branches to the phylogenetic tree, the value of the least-squares loss function Q was reduced from 1.660 to 1.410. The minimum value of the goodness-of-fit criterion Q_1 (Eq. 5) was reached at that point, having decreased from 0.00678 (without reticulation branches) to 0.00656 (with nine reticulation branches). Although the phylogenetic tree well represents the major thrust of the postglacial fish dispersal in the Québec peninsula (3.828 units of SS for 41 branches), the reticulation branches that were added (0.249 units of SS for 9 branches, or 0.0277 units per

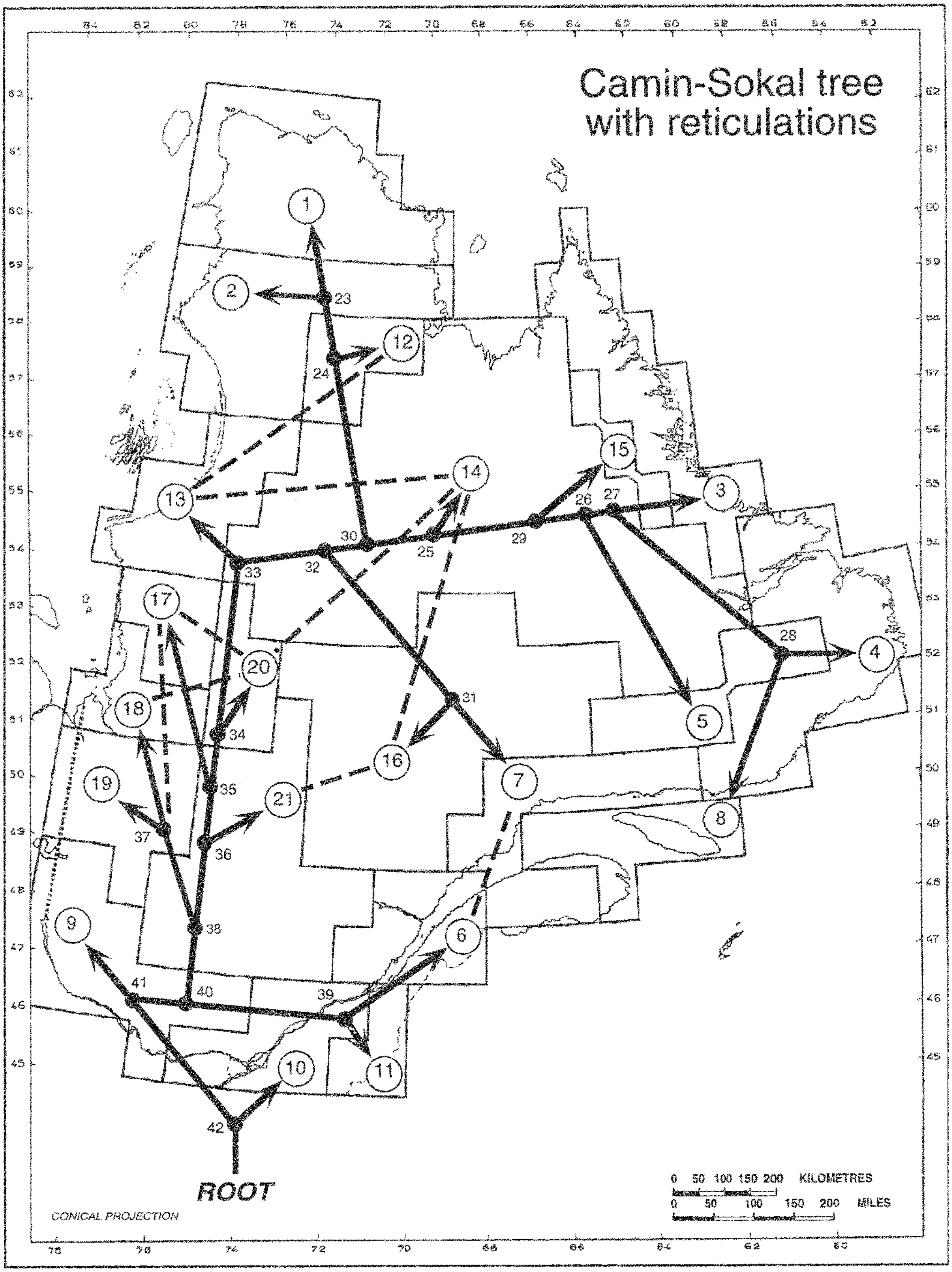

FIGURE 3. Map of the Québec peninsula divided into 21 biogeographic regions linked by a Camin–Sokal tree (thick solid lines) depicting possible postglacial dispersal routes for stenohaline freshwater fish species; reticulation branches (dashed lines) were added to that tree.

branch) represent a nonnegligible fraction of the similarity that was not represented by the phylogenetic tree. They correspond to fish migrations between neighboring territory units, using the river network, as explained in more detail by Legendre and Legendre (1984).

Example 2: Microgeographic Morphological Differentiation in Muskrats

Le Boulengé et al. (1996) studied the differentiation in cranial morphology of local populations of muskrats (*Ondatra zibethicus*) in a river network in southern Belgium near the French border. Muskrats are semiaquatic rodents introduced from North America into Europe in 1905 by Prince John of Bohemia as a valuable fur-bearing animal. In Belgium, they were released in nature in 1928 (Van Wijngaarden, 1955). Muskrats colonized the La Houille River network during the 1950s and exchanged genes among local populations until eradication during a trapping campaign carried out between October 1971 and February 1972. The part of the study that concerns us here includes nine local populations of muskrats found in a 50 km^2 area, inhabiting ponds scattered along the tributaries of River La Houille. The observed variability was attributed to a sociobiological mechanism called isolation by distance along corridors, a model for which Le Boulengé et al. (1996) have presented a detailed behavioral justification. They statistically tested 10 predictions originating from this model and involving morphological and geographic distances among the local populations of muskrats. The Mahalanobis distance was used to quantify the morphological distances among local populations, based on 10 age-adjusted linear measurements on skulls (mandible and cranium) (Table 4).

The river network (Fig. 4) is taken to represent an unrooted tree-like network corresponding to the main dispersal routes for muskrats among local populations. We used the age-adjusted Mahalanobis distance matrix among the nine populations to fit distances to the branches of the tree (Table 5).

Reticulations branches were added to the tree. Table 6 reports the contribution of each new reticulation branch to the reduction of the least-squares criterion Q (right-hand column). The minimum of the goodness-of-fit function Q_2 (Eq. 6) indicates that only the first four reticulation branches should be added (Fig. 5). The total sum of squares in the Mahalanobis distance matrix was 18.72 before fitting the tree. After fitting the tree, the residual SS was 3.420; that is, the 15 branches of the tree (Table 5) together explained 15.300 units of SS, or 1.02 unit per branch. Compare this result to the contributions to SS of the four first reticulation branches: 0.800, 0.438, 0.162, and 0.122 (Table 6). Those values are not negligible: the four reticulation branches help in an important way to explain the morphological distances among the local populations of muskrats.

Although the local populations were in general isolated from one another by forested areas crossed by small swift-water creeks, some (Zones M and Z on the one hand, and M and C on the other) were separated from each other by only short strips of swampy area; muskrats could readily move between these zones. For that reason, local populations M and Z were not included in the analysis of Le Boulengé et al. (1996) but they are included in the present study because they offer opportunities for reticulation branches. The headwaters of Zones N and O are also very close to each other, although the intervening area is forested. Three of the four reticulation branches proposed by the algorithm are between these neighbor areas: Zone M and node 10 (adjacent to Zone C), M and Z, and N and O. Gene exchange between these areas was expected but could not be expressed correctly by a nonreticulated network model. A fourth reticulation branch is identified between Zones J and N; it is not explained by geographic proximity and it thus presumably results from preferential migration along the river network, founder effect, genetic drift, or other chance mechanism.

Example 3: Phylogenetic Analysis of Aphelandra *and Hybrids*

In 1992, McDade published an analysis of a unique data set consisting of 50 morphological characters (coded into two to six states) measured over 12 species of Central American *Aphelandra* (perennial plants of the Acanthus family) and a series of 17 hybrids of known parental origin. The characters are described in McDade (1990: Appendix B).

We tried to reproduce the trees presented by McDade (1992), using the methodology described in that paper, but failed, perhaps

TABLE 4. Lower triangular Mahalanobis distance matrix for nine local populations of muskrats of the La Houille River basin, based on 10 age-adjusted linear measurements taken on skulls, for a total of 144 individuals (Le Boulengé et al., 1996: Table 1). This distance matrix is not reported in full in that paper; it was provided to us by Eric Le Boulengé, Université Catholique de Louvain, Belgium, whom we gratefully acknowledge.

Population zones	C	E	J	L	M	N	O	T	Z
C	0.0000								
E	2.1380	0.0000							
J	2.2713	2.9579	0.0000						
L	1.7135	2.3927	1.7772	0.0000					
M	1.5460	1.9818	2.4575	1.0125	0.0000				
N	2.6979	3.3566	1.9900	1.8520	2.6954	0.0000			
O	2.9985	3.6848	3.4484	2.4272	2.6816	2.3108	0.0000		
T	2.3859	2.3169	2.4666	1.4545	1.7581	2.2105	2.5041	0.0000	
Z	2.3107	2.3648	1.8086	1.6609	2.0516	2.2954	3.4301	2.0413	0.0000

because the data table has been updated by McDade since her 1992 paper; that certainly was the case for the ancestral states, which were stated for all characters in McDade (1990: Appendix B), whereas the data table she sent us contained six unknown states for the hypothesized ancestor. We proceeded as follows: First, a matrix of simple-matching similarity coefficients (S) was computed among the species subjected to the analysis and was transformed into a distance matrix by using the transformation $D = (1 - S)^{0.5}$.

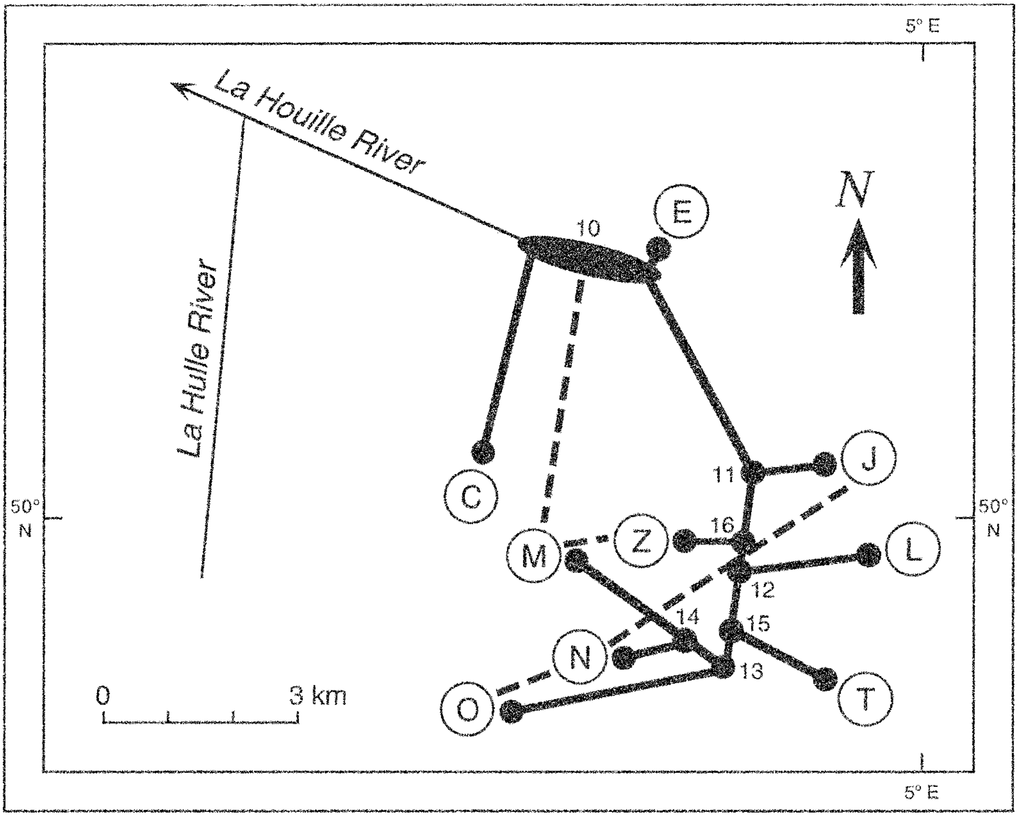

FIGURE 4. Schematic representation of the upper La Houille River network in Belgium, showing the nine muskrat population zones (terminal nodes identified by letters), the inner nodes (numbers), the tree branches (thick solid lines) and the reticulation branches (thick dashed lines). Node number 10, which sits on a river segment, is represented by an ellipse. Geographic coordinates of the area are given near the border of the map.

TABLE 5. List of branches and their lengths in the phylogenetic tree (represented by solid lines in Fig. 4).

Branch	Branch length	Branch	Branch length
10–E	1.041	12–16	0.242
11–J	0.940	12–15	0.000
12–L	0.480	15–T	0.743
13–O	1.851	10–C	1.169
14–M	0.650	10–11	0.051
13–15	0.093	11–16	0.044
13–14	0.000	16–Z	0.624
14–N	0.814		

The missing values in the data—six in the Ancestor and nine in each of the GOxLE and PAxLE hybrids—were handled by pairwise deletion during calculation of the similarity coefficient. We then reconstructed a phylogenetic tree, using the neighbor-joining method (Saitou and Nei, 1987), and added reticulation branches to the tree as described in the present paper. Addition of reticulation branches stopped when criterion Q_1 (Eq. 5) reached its minimum value.

The reconstructed tree (Fig. 6a) was fairly similar to the tree reported by McDade (1992: Fig. 1), the three main clades being identical. The total sum of squares in the distance matrix was 2.340 before fitting the tree, whereas after fitting the tree, the residual SS was 0.121; this result means that the 23 edges of the tree together explained 2.219 units of SS, or 0.096 unit per branch on average. In comparison, the five reticulation branches added to the tree explained together 0.048 units of SS—from 0.016 for the first reticulation branch (the 18–STOR link) to 0.004 for the fifth reticulation branch (the

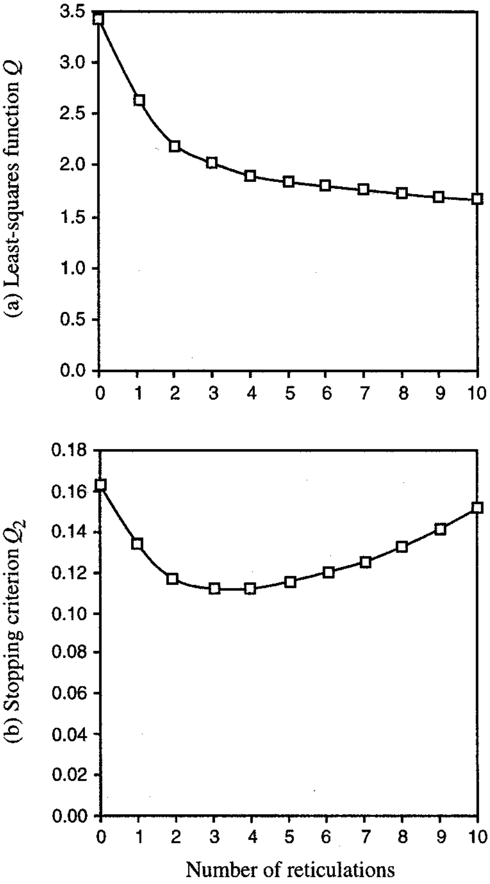

FIGURE 5. Behavior of (a) the least-squares function Q and (b) the goodness-of-fit criterion Q_2 for the first 10 iterations of the reticulogram reconstruction algorithm applied to the Mahalanobis distance matrix from Table 3. Abscissa: number of iterations of the algorithm. Zero corresponds to the phylogenetic tree before reticulation edges were added. The minimum value of Q_2 was reached at iteration 4.

TABLE 6. List of new branches with their lengths for the first 10 iterations of the algorithm. Q is the least-squares loss function, $Q_2 = Q/[n(n-1)/2] - N$ is the goodness-of-fit function; the minimum is reached at iteration 4 (bold). The right-hand column gives the contribution of each reticulation branch to the minimization of the sum of squares Q. The first four reticulation branches are represented by dashed lines in Figure 4.

Iteration	Branch	Branch length	Q	Q_2	Contribution to SS (Q)
0			3.420	0.1629	
1	N–O	1.770	2.620	0.1310	0.800
2	M–10	0.678	2.182	0.1148	0.438
3	J–N	1.729	2.020	0.1122	0.162
4	**M–Z**	**1.048**	**1.898**	**0.1117**	**0.122**
5	O–T	2.466	1.850	0.1156	0.048
6	C–J	1.948	1.805	0.1203	0.045
7	L–Z	1.137	1.761	0.1258	0.044
8	E–T	1.933	1.726	0.1328	0.035
9	L–T	1.067	1.702	0.1418	0.024
10	E–Z	1.608	1.679	0.1526	0.023

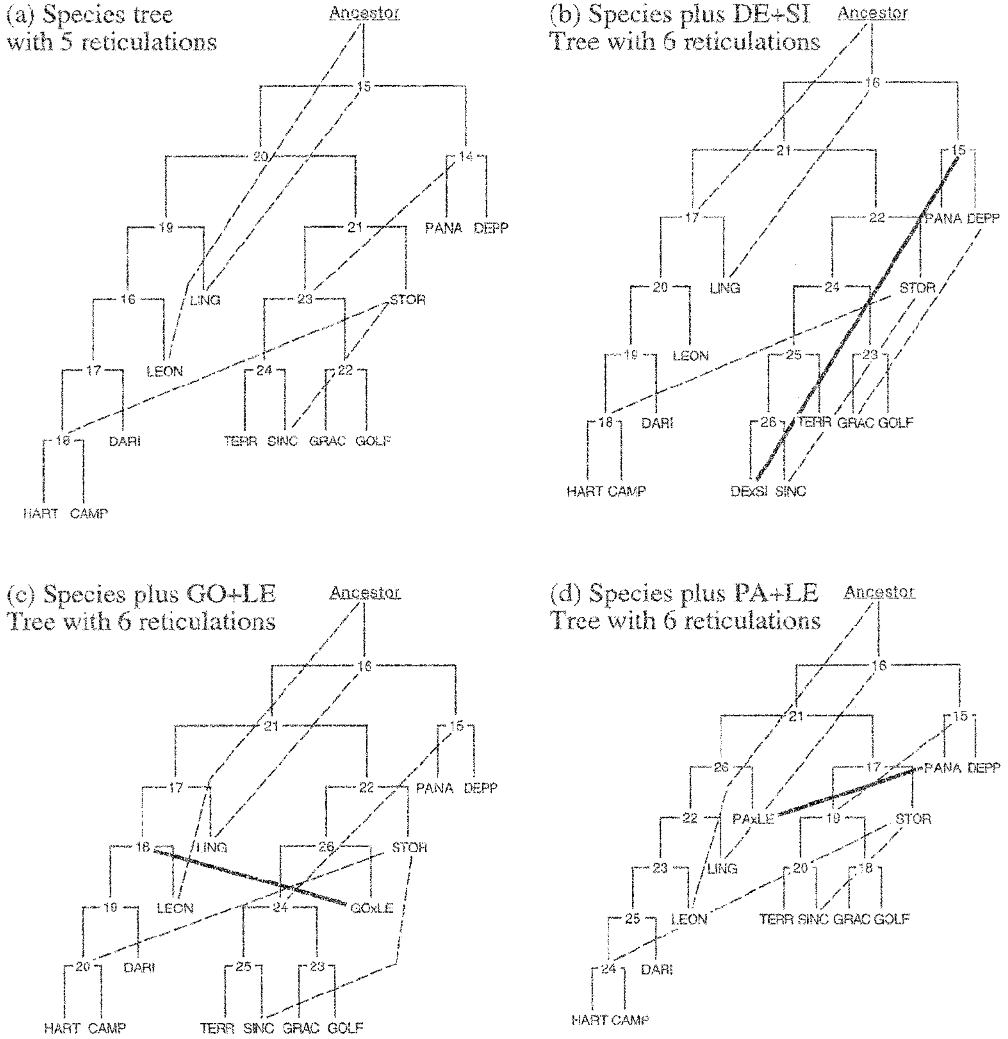

FIGURE 6. Reticulation analysis of McDade's *Aphelandra* data (a) for the 12 *Aphelandra* species and (b)–(d) for the same species plus one hybrid in each case (indicated in bold). Dashed lines: reticulation branches added to the tree (bold dashed lines: reticulation branches connecting to hybrids). Branches are not drawn to be proportional to the trees' branch lengths. Species name abbreviations: CAMP = *Aphelandra campanensis*; DARI = *A. darienensis*; DEPP = *A. deppeana* (formerly called *A. scabra*, and abbreviated SC in McDade's 1992 paper); GOLF = *A. golfodulcensis*; GRAC = *A. gracilis*; HART = *A. hartwegiana*; LING = *A. lingua-bovis*; LEON = *A. leonardii*; PANA = *A. panamensis*; SINC = *A. sinclairiana*; STOR = *A. storkii*; TERR = *A. terryae*; Ancestor = hypothesized ancestor. For hybrids, the abbreviations are constructed as follows: The first two letters of the ovulate parent's species name are followed by the first two letters of the staminate parent's name. For example, DExSI = DEPP (ovulate) × SINC (staminate).

SINC–STOR link)—or 0.010 units of SS per reticulation branch on average. By adding five reticulation branches to the phylogenetic tree, the value of the least-squares loss function Q was reduced from 0.121 to 0.073, corresponding to a 40% reduction of the residual SS of the tree. The least value of the goodness-of-fit criterion Q_1 was reached at that point, having decreased from 0.00634 (without reticulation branches) to 0.00541 (with five reticulation branches).

The data set of 12 species was reanalyzed together with the DExSI hybrid (Fig. 6b). The ovulate parent species, SINC, was the sister-group of the hybrid in the tree. The staminate parent, DEPP, was linked to the hybrid by a reticulation branch, indicated in bold, which actually connected to node 15, containing DEPP and PANA. The five reticulation branches present in Figure 6a were also found in Figure 6b, with little change. The Ancestor–LEON reticulation branch was moved to a higher node in Figure 6b, whereas the branch joining nodes 14 and 23 was moved to a lower position in the tree.

The GOxLE hybrid is found near the ovulate parent GOLF in Figure 6c and is linked to node 18 by a reticulation branch to which the staminate parent LEON is connected. The five other reticulation branches are identical to those in Figure 6a.

The PAxLE hybrid is found near the staminate parent LEON in Figure 6d, and it is linked to the ovulate parent PANA by a reticulation branch. The five other reticulation branches are identical to those in Figure 6a.

This example shows that hybrids may be identified by reticulation analysis. In each of these examples, the tree reconstruction method placed the hybrid near one of the parents in the tree, and the reticulation analysis linked it to the other parent by a reticulation branch. The other reticulation branches, which were present in all trees (without and with hybrids), display other features of the similarity, possibly homoplasy. Their positions were not much affected by the presence of hybrids in the analysis.

Discussion

We propose a new method for reconstructing reticulation networks from empirical distance matrices. We recommend its use in phylogenetic problems in which researchers suspect hybridization or wish to represent homoplasy in the data, and in biogeographic or microevolutionary problems in which a network is sought to represent the data. Starting with a phylogenetic tree, which provides an initial fit for a distance matrix, the algorithm improves on the tree solution by adding reticulation branches to the growing network. The method uses least squares or weighted least squares, placing optimally, during each iteration, a new reticulation branch onto the reticulogram under construction. The network inferred from a distance matrix of size $n \times n$ may have from $n + 1$ to $2n - 2$ nodes, depending on the topology of the phylogenetic tree reconstructed in the first part of the method, and from n to $(2n - 2)(2n - 3)/2$ branches in total. Such a structure, which comprises intermediate nodes, should provide a better fit to the distance matrix than will a phylogenetic tree. Note that each reticulation branch added by our algorithm represents a contradictory signal for which the phylogenetic tree reconstruction method has had to find compromises. The greater the number of extra branches added to a tree during reticulation analysis, the more conflicting features the original tree encompassed.

In different biogeographic or phylogenetic contexts, new reticulation branches may have different interpretations. In example 1, we showed that reticulation branches can represent migration routes among territory units that cannot be depicted by a phylogenetic tree. The reticulation network allowed us to link adjacent geographic areas possessing similar species assemblages and formulate new hypotheses about species migration routes. In example 3, reticulation branches linked hybrids to parent species, which happened to be known with certainty in that case.

Reticulograms can be of great interest in the study of bacterial evolution through lateral gene transfer, hybridization of eukaryotes, microevolution, and homoplasy, examples of which we give in a separate paper (Makarenkov and Legendre, in prep.). In a special journal section dedicated to reticulate evolution, Lapointe (2000), Legendre (2000b,c), Rohlf (2000), Smouse (2000), and Sneath (2000) described several applications for reticulation networks. Testing our method in each of these fields and providing plausible interpretations for reticulation branches should be of great interest for the development of evolutionary theory; we are making our computer program available to researchers to encourage them to carry out such studies. In each case, the interpretation should be based on the fact that any new reticulation branch linking two particular species indicates that the connected species are more closely related, or more similar for some other reason, than can be represented by a (bi)furcating phylogenetic tree. In some cases, reticulation branches may depict homoplasy (an example is given in Makarenkov and Legendre, 2000). Or they may represent possible hybridization or mutation events that occurred during evolution, or suggest that connected species could have a common ancestor. Examples of these cases remain to be explored.

A Monte Carlo study supported the superiority of reticulograms, using the criteria described in this paper, over phylogenetic trees for fitting reticulated data. The algorithm proposed in this paper remains, however, a heuristic strategy to approximate a distance matrix by a reticulogram with a fixed number of nodes; it does not guarantee an optimal solution. The resulting reticulogram depends heavily on the phylogenetic tree from which the reticulation network reconstruction algorithm starts; different initial trees may lead to different sets of reticulation branches. The present algorithm could be improved by rearranging the reticulogram topology during each iteration or by removing and replacing some branches. This will be the subject of another paper.

The heuristic algorithm for reconstructing reticulation networks described in this paper has been included in the T-REX package developed by Makarenkov and Casgrain (2000) (see Makarenkov, 2001). In addition to reticulation analysis, the T-REX program includes some popular phylogenetic tree-fitting algorithms: ADDTREE of Sattath and Tversky (1977), neighbor-joining of Saitou and Nei (1987), unweighted neighbor-joining of Gascuel (1997), the method of weighted least squares of Makarenkov and Leclerc (1999), and others. It is available as freeware to the research community at URL <http : //www.fas.umontreal.ca/biol/casgrain/en/labo/t-rex/>. T-REX allows users to visualize tree and reticulogram structures but does not map individual characters on trees or reticulate phylogenies.

Acknowledgments

We are grateful to Philippe Casgrain for his contribution to programming the T-REX package. Through this package, the new method is now available to the scientific community. We are also grateful to Eric Le Boulengé from Université Catholique de Louvain, Belgium, who provided the Mahalanobis distance matrix reported in Table 4; to Lucinda McDade, who kindly sent us her data table for the *Aphelandra* example and provided explanations about the data; and to Associate Editor Roberta Mason-Gamer as well as two anonymous reviewers for stimulating comments and discussion. This research was supported by NSERC grant number OGP7738 to P.L.

References

Alroy, J. 1995. Continuous track analysis: A new phylogenic and biogeographic method. Syst. Biol. 44: 152–178.

Bandelt, H.-J. 1995. Combination of data in phylogenetic analysis. Plant Syst. and Evol. Suppl. 9:355–361.

Bandelt, H.-J., and A. W. M. Dress. 1989. Weak hierarchies associated with similarity measures—A phylogenetic clustering technique. Bull. Math. Biol. 51: 133–166.

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

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

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

Bryant, D., and P. Waddell. 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.

Camin, J. H., and R. R. Sokal. 1965. A method for deducing branching sequences in phylogeny. Evolution 19:311–326.

Day, W. H. E. 1987. Computational complexity of inferring phylogenies from distance matrices. Bull. Math. Biol. 49:461–467.

Day, W. H. E. 1996. Complexity theory: An introduction for practitioners of classification. Pages 199–233 *in* Clustering and classification (P. Arabie, L. J. Hubert, and G. De Soete, eds.). World Scientific Publishers, River Branch, New Jersey.

De Soete, G., and J. D. Carroll. 1996. Tree and other network models for representing data. Pages 157–197 *in* Clustering and classification (P. Arabie, L. J. Hubert, and G. De Soete, eds.). World Scientific Publishers, River Branch, New Jersey.

Diday, E., and Bertrand, P. 1986. An extension of hierarchical clustering: The pyramidal representation. Pages 411–424 *in* Pattern recognition in practice (E. S. Gelsema and L. N. Kanal, eds.). North-Holland, Amsterdam.

Dress, A., D. Huson, and V. Moulton. 1996. Analyzing and visualizing sequence and distance data using SplitsTree. Discrete Appl. Math. 71:95–109.

Feger, H., and W. Bien. 1982. Network unfolding. Social Networks 4:257–283.

Feger, H., and U. Droge. 1984. Ordinale netzerkskalierung [Ordinal network scaling]. Kölner Zt. Soziol. Sozialpsychol. 3:417–423.

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

Gascuel, O. 1997. Concerning the NJ algorithm and its unweighted version, UNJ. Pages 149–170 *in* Mathematical hierarchies and biology. DIMACS Series in Discrete Mathematics and Theoretical Computer Science (B. Mirkin, F. R. McMorris, F. Roberts, and A. Rzhetsky, eds.). American Mathematical Society. Providence, Rhode Island.

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

Jakobsen, B., and S. Easteal. 1996. A program for calculating and displaying compatibility matrices as an aid in determining reticulate evolution in molecular sequences. CABIOS 12:291–295.

Klauer, K. C. 1988. Representing proximities by ordinal networks. Pages 473–477 *in* Classification and related methods of data analysis (H. H. Bock, ed.). North-Holland, Amsterdam.

Klauer, K. C. 1989. Ordinal network representation: Representing proximities by graphs. Psychometrika 54:737–750.

Klauer, K. C., and J. D. Carroll. 1989. A mathematical programming approach for fitting general graphs. J. Classif. 6:247–270.

Klauer, K. C., and J. D. Carroll. 1995. Network models for scaling proximity data. Pages 319–342 *in* Geometric representations of perceptual phenomena (R. D. Luce, M. D'Zmura, D. D. Hoffman, G. Iverson, and R. K. Romney, eds.). Erlbaum, Hillsdale, New Jersey.

Lapointe, F.-J. 2000. How to account for reticulation events in phylogenetic analysis: A comparison of distance-based methods. J. Classif. 17:175–184.

Lapointe, F.-J., and P. Legendre. 1992. A statistical framework to test the consensus among additive trees (cladograms). Syst. Biol. 41:158–171.

Le Boulengé, É., P. Legendre, C. De Le Court, P. Le Boulengé-Nguyen, and M. Languy. 1996. Microgeographic morphological differentiation in muskrats. J. Mammal. 77:684–701.

Legendre, P. 1984. Report on Seventeenth International Numerical Taxonomy Conference. Syst. Zool. 33:117–121.

Legendre, P. 1986. Reconstructing biogeographic history using phylogenetic-tree analysis of community structure. Syst. Zool. 35:68–80.

Legendre, P. (Ed.). 2000a. Special section on reticulate evolution. J. Classif. 17:153–195.

Legendre, P. 2000b. Reticulate evolution: From bacteria to philosopher. J. Classif. 17:153–157.

Legendre, P. 2000c. Biological applications of reticulation analysis. J. Classif. 17:191–195.

Legendre, P., and L. Legendre. 1998. Numerical ecology, 2nd English edition. Elsevier Science BV, Amsterdam.

Legendre, P., and V. Legendre. 1984. Postglacial dispersal of freshwater fishes in the Québec peninsula. Can. J. Fish. Aquat. Sci. 41:1781–1802.

Legendre, V. 1942. Redécouverte après un siècle et reclassification d'une espèce de catostomidé. Nat. Can. 69:227–233.

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

Makarenkov, V., and P. Casgrain. 2000. T-REX package of application programs for tree and reticulogram reconstruction. Univ. de Montréal, Département de Sciences Biologiques, Montréal.

Makarenkov, V., and B. Leclerc. 1999. An algorithm for the fitting of a tree metric according to a weighted least-squares criterion. J. Classif. 16:3–27.

Makarenkov, V., and P. Legendre, 2000. Improving the additive tree representation of a given dissimilarity matrix using reticulations. Pages 35–40 *in* Data analysis, classification and related methods (H. A. L. Kiers, J.-P. Rasson, P. J. F. Groenen, and M. Schader, eds.). Springer, Namur, Belgium.

McDade, L. A. 1990. Hybrids and phylogenetic systematics I. Patterns of character expression in hybrids and their implications for cladistic analysis. Evolution 44:1685–1700.

McDade, L. A. 1992. Hybrids and phylogenetic systematics II. The impact of hybrids on cladistic analysis. Evolution 46:1329–1346.

Orth, B. 1988. Representing similarities by distance graphs: Monotonic network analysis MONA. Pages 489–494 *in* Classification and related methods of data analysis (H. H. Bock, ed.). North-Holland, Amsterdam.

Pruzansky, S., A. Tversky, and J. D. Carroll. 1982. Spatial versus tree representations of proximity data. Psychometrika 47:3–19.

Rohlf, F. J. 2000. Phylogenetic models and reticulations. J. Classif. 17:185–189.

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

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

Smouse, P. E. 2000. Reticulation inside the species boundary. J. Classif. 17:165–173.

Sneath, P. H. A. 1975. Cladistic representation of reticulate evolution. Syst. Zool. 24:360–368.

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

Swofford, D. L., G. J. Olsen, P. J. Waddell, and D. M. Hillis. 1996. Phylogenetic inference. Pages 407–514 *in* Molecular systematics (D. M. Hillis, C. Moritz, and B. K. Mable, eds.). Sinauer Associates, Sunderland, Massachusetts.

Van Wijngaarden, A. 1955. De bestrijding van de muskusrat in Nederland. Vakbl. Biol. 35:68–72.

Wanntorp, H.-E. 1983. Reticulated cladograms and the identification of hybrid taxa. Pages 81–99 *in* Advances in cladistics (N. I. Platnick and V. A. Funk, eds.). Columbia Univ. Press, New York.

Zaretskii, K. 1965. Construction of a tree on the basis of a set of distances between its leaves. Usp. Mat. Nauk 20:90–92.

Received 15 March 2001; accepted 6 December 2001
Associate Editor: Roberta Mason-Gamer