Journal of Classification 17:175-184 (2000)
DOI: 10.1007/s003570000016

Special Section on Reticulate Evolution

How to Account for Reticulation Events in Phylogenetic Analysis: A Comparison of Distance-Based Methods

François-Joseph Lapointe

Université de Montréal

Abstract: This paper presents a review of mathematical techniques capable of representing reticulate events in phylogenetics. Two families of methods are identified; they relax either the ultrametric inequality defining dendrograms or the four-point condition defining additive trees. Pyramids and weak hierarchies are techniques developed to fit dendrograms with overlapping clusters. Splitsgraphs and reticulograms are extensions of the additive tree model; they allow one to fit a dissimilarity matrix using a graph containing reticulations. The four methods are applied to a data set; the results are compared and discussed in a phylogenetic setting.

Keywords: Additive tree; Dendrogram; General graph; Pyramids; Reticulogram; Splitsgraph; Weak hierarchy.

I am grateful to Guy Cucumel, Vladimir Makarenkov, David Bryant, and Andreas Dress for helping me understand some of the mathematical aspects of the methods presented in this paper, to Pierre Legendre for inviting me to participate in this Special Section, to Jim Rohlf for encouraging me to write this paper, and to Peter Sneath and Peter Smouse for constructive comments on a previous version. This research was supported by grants from the Natural Sciences and Engineering Research Council of Canada and the Fonds pour la Formation de Chercheurs et l'Aide à la Recherche du Québec.

Author's Address: François-Joseph Lapointe, Département de sciences biologiques, Université de Montréal, C.P. 6128, succursale Centre-ville, Montréal, Québec H3C 3J7, Canada; e-mail: francois-joseph.lapointe@umontreal.ca

1. Introduction

The basic graph-theoretic model used in phylogenetic analysis is that of a tree. Sometimes, trees are rooted to represent ancestor-descendent relationships among their nodes, and sometimes, their branches are weighted to represent the amount of evolutionary change along those branches. Two different types of weighted trees are commonly used to depict evolutionary relationships among species. *Dendrograms* are used to represent rooted weighted trees in which all terminal nodes are equidistant from the root, whereas *additive trees* are used to represent unrooted weighted trees; additive trees can also be rooted by selecting one of the nodes to form the root of the tree. The distinction is important in evolutionary biology because dendrograms represent trees that satisfy the molecular clock hypothesis stating that all lineages evolved at the same rate (Figure 1a). This assumption is not always made for additive trees (Figure 1d). For the purpose of the present paper, it suffices to say that dendrograms satisfy the well-known ultrametric inequality (Hartigan 1967) and that additive trees satisfy the more general four-point condition (Buneman 1974); a more detailed presentation of dendrograms and additive trees is found in Lapointe and Legendre (1991).

A tree is not always a suitable graphical representation of the evolutionary relationships among species. In fact, it is not uncommon for species to exchange genetic material laterally instead of vertically (along the branches of the tree). These so-called reticulation events violate the branching evolutionary model by introducing cycles in a graph and causing conflicting signals in the data. Other representations must be used to depict such evolutionary phenomena, which cannot adequately be represented in the form of trees. Four such *reticulistic* techniques are described below and are applied to the same data set for comparison.

2. Pyramids

Pyramids, introduced by Diday and Bertrand (1986), are a generalization of the hierarchical clustering framework. Whereas a dendrogram can be defined as a nested set of nonoverlapping clusters (Figure 1a), pyramids represent a set of clusters that may overlap without necessarily being nested (Figure 1b). For any given pair of clusters C and D in a dendrogram **H** that have a nonempty intersection, either C is contained in D, or D is contained in C. In Figure 1a, for example, the cluster {*Pan paniscus*, *Pan troglodytes*} is contained in the cluster {*Homo sapiens*, *Pan paniscus*, *Pan troglodytes*} of **H**. In the case of a pyramidal graph **P**, the intersection of two clusters C and D that have a nonempty intersection is always a cluster of **P**. In Figure 1b, for example, the intersection of the clusters {*Pan paniscus*, *Pan troglodytes*} and {*Homo sapiens*, *Pan troglodytes*} is the singleton {*Pan troglodytes*} of **P**.

A dissimilarity matrix **D** is said to be pyramidal iff **D** is also a Robinsonian matrix (Table 1c, upper triangular). This property means that, for any triplet *i, j, k*, from an *ordered* set of species, the dissimilarity value d_{ik} must be larger than or equal to the maximum of d_{ij} and d_{jk}. Interestingly, an ultrametric matrix **U** can always be permuted to form a Robinsonian matrix, so that a dendrogram actually represents a special type of pyramids with at most n–1 different clusters. Just like dendrograms, pyramids can be obtained by agglomerative algorithms. In Figure 1b, the pyramidal representation of the dissimilarities presented in Table 1a was obtained from the equivalent of the complete linkage algorithm: two clusters are joined at a given height if they satisfy the clustering rule and have not been aggregated *twice* before; in the case of dendrograms, two clusters are joined if they satisfy the clustering rule and have not been aggregated *once* before. By allowing species to be included in overlapping clusters, pyramids can thus be used to depict reticulation events in a set of species that can be ordered in a Robinsonian matrix. A program to compute pyramids is available at the following WWWeb address: <http://genome.genetique.uvsq.fr/ Pyramids/>.

3. Weak Hierarchies

Weak hierarchies have been proposed by Bandelt and Dress (1989) to fit dendrograms with a few additional nonnested clusters (i.e., reticulations). In short, the method proceeds by creating weak clusters of species, as opposed to the so-called strong clusters found in dendrograms. From a similarity matrix **S**, a weak cluster C is formed if any two species *i* and *j* that belong to C are more similar to each other than any other species *k* outside of C is similar to at least one of *i* and *j* (Bandelt and Dress 1989); that is, s_{ij} must be larger than the minimum of s_{ik} and s_{kj} for every species *k* which is not a member of C (for strong clusters, s_{ij} must be larger than the maximum of s_{ik} and s_{kj}). Using a set-theoretic point of view, a weak hierarchy **W** is obtained if the intersection of any three (strong or weak) clusters C, D, and E of **W** is equal to one of the binary intersections $C \cap D$, $C \cap E$, or $D \cap E$. For example, one can check that the similarity s_{ij} between *Homo sapiens* and *Pan troglodytes* in **S** (where $s_{ij} = 1 - d_{ij}$; Table 1a) is larger than the lesser of the similarities s_{ik} and s_{kj} between any other species *k* and either *H. sapiens* or *P. troglodytes*; as a consequence, the pair {*Homo sapiens*, *Pan troglodytes*} represents a weak cluster of **W**. Then, because

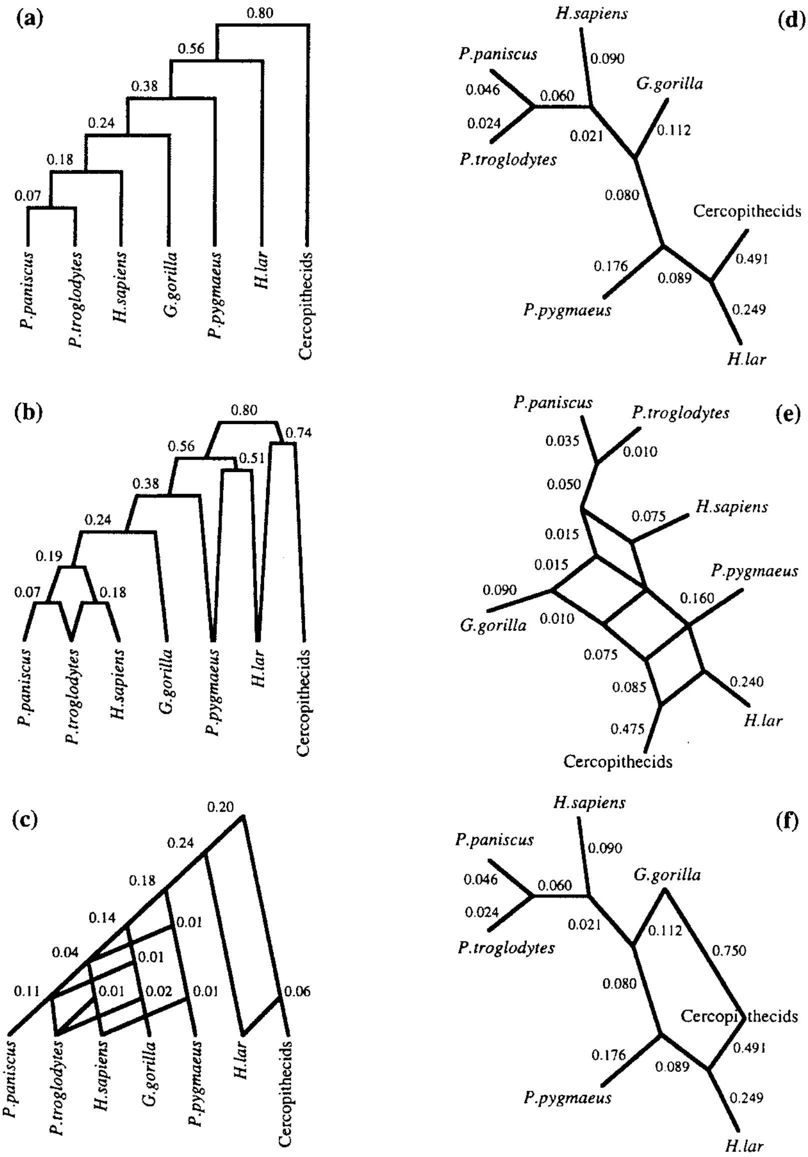

Figure 1. Different representations of the dissimilarity matrix of Table 1a. (a) A complete linkage dendrogram with clustering levels. The distance between two species is given by the height of the lowest cluster that includes these species. The corresponding ultrametric distances are presented in Table 1b (upper triangular). (b) Complete linkage pyramids with clustering levels. The distance between two species is given by the height of the lowest cluster that includes these species. The corresponding (Robinsonian) pyramidal distances are presented in Table 1c (upper triangular). (c) A weak hierarchy (*caption continues on next page*)

the intersection of the clusters {*H. sapiens*, *P. paniscus*, *P. troglodytes*, *G. gorilla*}, {*H. sapiens*, *P. paniscus*, *P. troglodytes*} and {*H. sapiens*, *P. troglodytes*} in the weak hierarchy derived from Table 1a is equal to at least one of the three binary intersections, these four species form a weak hierarchy (see Figure 1c).

A weak hierarchy is an extension of a dendrogram and represents all weak and strong clusters. Consequently, any dendrogram is a weak hierarchy, whereas pyramids are nothing but weak hierarchies with the additional property that a linear order of the species can be defined such that every cluster is an interval relative to that order. Using the clusters of a weak hierarchy, one can compose a similarity matrix additively (Table 1c, lower triangular; where $d_{ij} = 1 - s_{ij}$) by attaching a weight to each cluster and letting the similarity of a pair of species i and j be the sum of the weights of all the clusters (weak or strong) containing the pair $\{i, j\}$; see the algorithm in Bandelt and Dress (1989). Furthermore, given the weighted weak hierarchy, one can reconstruct all of its clusters as well as their respective weights from the associated similarity matrix. A complete linkage type of algorithm has been developed by Bandelt and Dress (1989) to approximate a similarity matrix **S** by a weak hierarchy (see Figure 1c). A computer program to compute weak hierarchies can be obtained by writing to Professor H.-J. Bandelt: Mathematisches Seminar, Universität Hamburg, Bundesstrasse 55, D-20146 Hamburg, Germany.

4. Splitsgraph

As in the case of dendrograms, reticulations are not allowed in additive trees (see Figure 1d). To produce unrooted phylogenies in which

Figure 1 (*continued*) obtained by a complete linkage-type method applied to the matrix presented in Table 1a, with corresponding weights attached to the clusters. The similarity between two species is computed as the sum of the weights of all the clusters that include these species. The corresponding distances are presented in Table 1c (lower triangular), where $d_{ij} = 1 - s_{ij}$. (d) An additive tree, with edge lengths, obtained by a least-squares algorithm. The distance between two species is computed as the sum of the edge lengths along the path connecting these species. The corresponding path-length distances are presented in Table 1b (lower triangular). (e) A splitsgraph representation with edge lengths; all parallel edges have equal lengths. The distance between two species is computed as the shortest path-length distance between these species over all possible paths. The corresponding path-length distances are presented in Table 1d (upper triangular). (f) A reticulogram, with edge lengths, obtained by adding reticulations onto the additive tree presented in Figure 1d. The distance between two species is computed as the shortest path-length distance between these species over all possible paths. The corresponding path-length distances are presented in Table 1d (lower triangular). For clarity, edge lengths in the figure are not represented proportional to their actual lengths.

Table 1

a: Initial dissimilarity matrix (modified from Bandelt and Dress 1989)

	1	2	3	4	5	6	7
1. *H. sapiens*	0.0000	0.1900	0.1800	0.2400	0.3600	0.5200	0.7700
2. *P. paniscus*	0.1900	0.0000	0.0700	0.2300	0.3700	0.5600	0.8000
3. *P. troglodytes*	0.1800	0.0700	0.0000	0.2100	0.3700	0.5100	0.7700
4. *G. gorilla*	0.2400	0.2300	0.2100	0.0000	0.3800	0.5400	0.7500
5. *P. pygmaeus*	0.3600	0.3700	0.3700	0.3800	0.0000	0.5100	0.7600
6. *H. lar*	0.5200	0.5600	0.5100	0.5400	0.5100	0.0000	0.7400
7. Cercopithecids	0.7700	0.8000	0.7700	0.7500	0.7600	0.7400	0.0000

b: Distances corresponding to the dendrogram of Figure 1a (upper triangular matrix) and the additive tree of Figure 1d (lower triangular)

	1	2	3	4	5	6	7
1. *H. sapiens*	0.0000	0.1800	0.1800	0.2400	0.3800	0.5600	0.8000
2. *P. paniscus*	0.1960	0.0000	0.0700	0.2400	0.3800	0.5600	0.8000
3. *P. troglodytes*	0.1739	0.0701	0.0000	0.2400	0.3800	0.5600	0.8000
4. *G. gorilla*	0.2233	0.2393	0.2173	0.0000	0.3800	0.5600	0.8000
5. *P. pygmaeus*	0.3672	0.3832	0.3612	0.3683	0.0000	0.5600	0.8000
6. *H. lar*	0.5287	0.5447	0.5227	0.5298	0.5140	0.0000	0.8000
7. Cercopithecids	0.7707	0.7867	0.7647	0.7719	0.7560	0.7400	0.0000

c: Distances corresponding to the pyramids of Figure 1b (upper triangular matrix[1]) and the weak hierarchy of Figure 1c (lower triangular[2])

	1	2	3	4	5	6	7
1. *H. sapiens*	0.0000	0.1900	0.1800	0.2400	0.3800	0.5600	0.8000
2. *P. paniscus*	0.1900	0.0000	0.0700	0.2400	0.3800	0.5600	0.8000
3. *P. troglodytes*	0.1800	0.0700	0.0000	0.2400	0.3800	0.5600	0.8000
4. *G. gorilla*	0.2400	0.2300	0.2100	0.0000	0.3800	0.5600	0.8000
5. *P. pygmaeus*	0.3600	0.3700	0.3700	0.3800	0.0000	0.5100	0.8000
6. *H. lar*	0.5600	0.5600	0.5600	0.5600	0.5600	0.0000	0.7400
7. Cercopithecids	0.8000	0.8000	0.8000	0.8000	0.8000	0.7400	0.0000

[1] The matrix is Robinsonian if the species are ordered as in the pyramids {2, 3, 1, 4, 5, 6, 7}.

[2] The distances were obtained by subtracting the similarity values from one: $d_{ij} = 1 - s_{ij}$.

d: Distances corresponding to the splitsgraph of Figure 1e (upper triangular matrix) and the reticulogram of Figure 1f (lower triangular)

	1	2	3	4	5	6	7
1. *H. sapiens*	0.0000	0.1700	0.1450	0.2050	0.3250	0.4900	0.7400
2. *P. paniscus*	0.1960	0.0000	0.0450	0.2050	0.3450	0.5100	0.7600
3. *P. troglodytes*	0.1739	0.0701	0.0000	0.1800	0.3200	0.4850	0.7350
4. *G. gorilla*	0.2233	0.2393	0.2173	0.0000	0.3500	0.5150	0.7350
5. *P. pygmaeus*	0.3672	0.3832	0.3612	0.3683	0.0000	0.4850	0.7350
6. *H. lar*	0.5287	0.5447	0.5227	0.5298	0.5140	0.0000	0.7300
7. Cercopithecids	0.7707	0.7867	0.7647	0.7500	0.7560	0.7400	0.0000

the species manifest reticulation, the splitsgraph method of Bandelt and Dress (1992) can be used. This technique relies on split decomposition, a procedure for decomposing distances canonically into a sum of simpler metrics. For each quadruplet of species *{i, j, k, l}*, the algorithm looks at the three possible tree topologies that may be used to split the four species in two groups of two (i.e., *ij/kl*, *ik/jl*, and *il/jk*), with their corresponding sum of distances (i.e., *ij* + *kl*, *ik* + *jl*, and *il* + *jk*). Instead of *selecting the most probable* topology (i.e., the one with the smallest distance sum) as the estimate of the relationships, Bandelt and Dress's (1992) method *excludes the most improbable* of the three topologies at each step. The global splits (computed over all possible quadruplets) which never realize the most improbable topologies are accepted and depicted as a "splitsgraph" (Dress, Huson, and Moulton 1996).

In contrast to additive trees in which any edge splits the tree into two connected subtrees, incompatible splits cannot always be depicted by a single edge but will give rise to a series of parallel edges of equal lengths; the length of these parallel edges represents the isolation index of a given split. Therefore, a splitsgraph is a representation, composed of parallelograms plus individual edges, providing a visual representation of the support for contradictory patterns in the data (see Figure 1e). Unlike additive trees, in which the path-length distance between two species *i* and *j* is computed by adding the edge lengths along the path between these species, path-lengths in a splitsgraph are the shortest lengths of all paths from species *i* to *j* (also corresponding to the sum of all weighted splits separating two species, Table 1d, upper triangular). For example, the path length between *Pan paniscus* and *Gorilla gorilla* in Figure 1e is 0.2050, instead of 0.2393 in the additive tree of Figure 1d. A splittability index can be used to indicate the fit of the weighted system of splits, depicted as a splitsgraph, to the original dissimilarities in **D**. A computer program to compute splitsgraphs is available at the following WWWeb address: <http://bibiserv.techfak.uni-bielefeld.de/splits/>.

5. Additive Tree with Reticulations

Recently, Makarenkov and Legendre (submitted) have proposed an algorithm to add reticulations onto an additive tree so as to maximize the fit between the data and a reticulogram, which is an evolutionary graph in which the data may be related nonuniquely to a common ancestor (Makarenkov and Legendre 2000). This graph is computed by gradually improving the approximation of the dissimilarities as extra edges are added to the graph. Contrary to the other methods, this technique uses an optimality criterion to determine the minimum number of reticulations required to reach a maximum fit to the data; a least-squares loss function computed as the sum of the squared differences between the original dissimilarities in **D** and the path-length distances **P** on the reticulogram is minimized. Because there is more than one way to compute the path lengths between two species *i* and *j*, the minimum path-length distance over all possible paths from *i* to *j* is recorded in **P** (see Table 1d, lower triangular). For instance, the path-length distance between *Gorilla gorilla* and the Cercopithecids in Figure 1f is the length of the reticulate edge connecting these two species (0.7500) rather than the sum of the edge lengths along the original and unique path found in the additive tree (0.7719 in Figure 1d).

Makarenkov and Legendre (submitted) described three stopping rules to determine the number of reticulations to be added to an additive tree. Criterion Q1 takes into account the value of the loss function as well as the number of degrees of freedom of the reticulogram under construction; two other criteria, Q2 and AIC, have also been proposed by those authors. A statistical procedure could possibly be implemented to assess the significance of individual reticulations, using the Q1 statistic, for a graph bearing *n* edges compared to one with *n*–1 edges. A program to compute additive trees and reticulograms is available at the WWWeb address <http://www.fas.umontreal.ca/biol/legendre/>.

6. Discussion

To produce reticulograms, it is difficult to select a single best method among the four described in this paper. Pyramids allow for overlapping clusters and can perfectly fit a dissimilarity matrix if there exists a permutation order of the species such that the dissimilarities are Robinsonian. In the case of weak hierarchies, an optimal collection of weighted weak clusters is sought to reconstruct a similarity measure that approximates the original similarities. Both methods should therefore be able to fit the dis/similarities better than a dendrogram without reticulations. When reticulograms based on extended additive trees are sought, the splitsgraph method, which detects incompatible splits in the data, can be used to obtain a graphical representation of a dissimilarity matrix. Allowing cycles in a graph produces in turn a better fit of the model to the data. One can also use the method proposed by Makarenkov and Legendre (2000) which seeks to improve the representation of a dissimilarity matrix by adding reticulations to a previously estimated additive tree.

In the example used throughout this paper, a cophenetic correlation of 0.99749 was obtained between the ultrametric matrix (Table 1b, upper triangular) and the input dissimilarity matrix (Table 1a), indicating a good fit of the data by a dendrogram. A slightly larger correlation of 0.99897 was found between the input data and the path-length distances associated with the additive tree (Table 1b, lower triangular). As expected, the correlations for all reticulistic methods were even larger. For the pyramidal distances (Table 1c, upper triangular) and the distances associated with the weak hierarchy (Table 1c, lower triangular), the correlations were respectively 0.99769 and 0.99754. Similarly, correlations of 0.99914 and 0.99922 were obtained for the splitsgraph (Table 1d, upper triangular) and the additive tree with one extra edge (Table 1d, lower triangular).

Interestingly, the various methods produced somewhat different results; the biological meaning of these representations is of great importance. Whereas overlap is only allowed among contiguous clusters of species in pyramids, weak hierarchies can be used to represent reticulations between distant species or clusters (see Bandelt and Dress 1989). Similarly, the extra edges fitted on a tree when using the Makarenkov and Legendre algorithm tend to join distant species, as shown by the various examples presented by these authors (Makarenkov and Legendre 2000, and submitted). In such cases, reticulations may simply represent incompatibilities in the data resulting from convergent evolution. Another option, allowing the detection of a larger number of incompatibilities, is the splitsgraph. However, since they create a series of multiple parallel edges, splitsgraphs may quickly be saturated with extra vertices and edges, making it difficult to display them as planar graphs (Dress, Huson, and Moulton 1996).

7. Conclusion

This paper presented four different but somewhat related approaches to account for reticulation events in phylogenetic analysis. This list is not exhaustive; other techniques are currently available and being developed to produce reticulograms from gene frequencies (Xu 2000), binary data (Smouse 1998), or multistate characters using median graphs (Bandelt, Forster, and Rohl 1999). There are also clustering methods that can produce overlapping clusters. It is worth mentioning that the split decomposition method (Bandelt and Dress 1992) can be applied in other contexts than with distance data. To produce a splitsgraph all one needs is a phylogenetic method (parsimony or maximum likelihood) to decide, for each quadruplet, which of the three topologies is the most inappropriate. Likewise, an evolutionary parsimony criterion could be used to modify the Makarenkov and Legendre approach. Instead of searching for a reticulogram minimizing a least-squares criterion, extra edges could be added to parsimonious trees obtained from standard algorithms so as to minimize the number of character-state changes on those trees. Whichever approach is selected, one should be aware that in spite of interesting mathematical properties, the different reticulistic methods will not necessarily produce biologically meaningful results. Model-based techniques should be developed to serve that purpose. On the other hand, simulation studies are badly needed to evaluate the relative performances of the extant competing methods. In addition, more comparative studies are required to determine the success rate of the different algorithms to recover known phylogenies that include species of reticulate origins like hybrids or allopolyploids (e.g., McDade 1997).

References

BANDELT, H. J., and DRESS, A. W. (1989), "Weak Hierarchies Associated With Similarity Measures: An Additive Clustering Technique," *Bulletin of Mathematical Biology, 51,* 133-166.

BANDELT, H. J., and DRESS, A. W. (1992), "Split Decomposition: A New and Useful Approach to Phylogenetic Analysis of Distance Data," *Molecular Phylogenetics and Evolution, 1,* 242-252.

BANDELT, H. J., FORSTER, P., and ROHL, A. (1999), "Median-Joining Networks for Inferring Intraspecific Phylogenies," *Molecular Biology and Evolution, 16,* 37-48.

BUNEMAN, P. (1974), "A Note on the Metric Properties of Trees," *Journal of Combinatorial Theory (B), 17,* 48-50.

DIDAY, E., and BERTRAND, P. (1986), "An Extension of Hierarchical Clustering: The Pyramidal Representation," in *Pattern Recognition in Practice*, Eds., E. S. Gelsema and L. N. Kanal, Amsterdam: North-Holland, 411-424

DRESS, A., HUSON, D., and MOULTON, V. (1996), "Analyzing and Visualizing Sequence and Distance Data Using SPLITSTREE," *Discrete Applied Mathematics, 71,* 95-109.

HARTIGAN, J. A. (1967), "Representation of Similarity Matrices by Trees," *Journal of the American Statistical Association, 62,* 1140-1158.

LAPOINTE, F.-J., and LEGENDRE, P. (1991), "The Generation of Random Ultrametric Matrices Representing Dendrograms," *Journal of Classification, 8,* 177-200.

MAKARENKOV, V., and LEGENDRE, P. (2000), "Improving the Additive Tree Representation of a Dissimilarity Matrix Using Reticulations," in *Data Analysis, Classification, and Related Methods,* Eds., H. A. L. Kiers, J.-P. Rasson, P. J. F. Groenen, and M. Schader, Berlin: Springer, 35-46.

MAKARENKOV, V., and LEGENDRE, P. "General Network Representation of a Dissimilarity Matrix: Adding Reticulations to an Additive Tree," (submitted).

McDADE, L. A. (1997), "Hybrids and Phylogenetic Systematics III. Comparison With Distance Methods," *Systematic Botany, 22,* 669-683.

SMOUSE, P. E. (1998), "To Tree or Not To Tree," *Molecular Ecology, 7,* 399-412.

XU, S. (2000), "Phylogenetic Analysis Under Reticulate Evolution," *Molecular Biology and Evolution, 17,* 897-907.