Improving the additive tree representation of a dissimilarity matrix using reticulations

Vladimir Makarenkov[1] and Pierre Legendre[2]

[1] Département de sciences biologiques, Université de Montréal,
C.P. 6128, succ. Centre-ville, Montréal, Québec H3C 3J7, Canada
and Institute of Control Sciences, 65 Profsoyuznaya, Moscow 117806, Russia
[2] Département de sciences biologiques, Université de Montréal,
C.P. 6128, succ. Centre-ville, Montréal, Québec H3C 3J7, Canada

Abstract. This paper addresses the problem of approximating a dissimilarity matrix by means of a reticulogram. A reticulogram represents an evolutionary structure in which the objects may be related in a non-unique way to a common ancestor. Dendrograms and additive (phylogenetic) trees are particular cases of reticulograms. The reticulogram is obtained by adding edges (reticulations) to an additive tree, gradually improving the approximation of the dissimilarity matrix. We constructed a reticulogram representing the evolution of 12 primates. The reticulogram not only improved the data approximation provided by the phylogenetic tree, but also depicted the homoplasy contained in the data, which cannot be expressed by a tree topology. The algorithm for reconstructing reticulograms is part of the T-Rex software package, available at URL <http://www.fas.umontreal.ca/BIOL/legendre>.

1 Introduction

Several algorithms have been proposed for the representation of empirical dissimilarity data using a general network where the objects are represented by the nodes of a valued graph whose minimum path-length distances are associated with the dissimilarities (Feger and Bien 1982; Orth 1989; Klauer and Carroll 1989). An expanding tree structure based on weak clusters has also been proposed by Bandelt and Dress (1989) leading to a weak hierarchy for an empirical similarity matrix. Bandelt and Dress (1992) and Bandelt (1995) resumed investigation of weak clusters and proposed the method of split decomposition.

We outline the main features of a reticulogram reconstruction algorithm offering another way of modelling a dissimilarity matrix by means of a network. Our representation uses a topology called a *reticulogram* which includes the vertices associated with the objects in a set X as well as the intermediate nodes. A reticulogram can represent relationships among objects that may be related in a non-unique way to a common ancestor; such a structure cannot be represented by a tree. In a reticulogram, the distance between i and j is the *minimum-path-length distance* over the set of all paths linking i and j.

Inferring an additive tree from a dissimilarity matrix is a very well-studied issue in the literature. We launch the reticulogram reconstruction algorithm from an additive tree topology providing an initial fit for the dissimilarity matrix. The algorithm adds new edges or *reticulations* to a growing reticulogram, minimising the least-squares loss function computed as the sum of the quadratic differences between the original dissimilarities and the associated reticulogram estimates.

Reticulate patterns are found in nature in some phylogenetic problems. (1) In bacterial evolution, lateral gene transfer (LGT) produces reticulate evolution; LGT represents the mechanisms by which bacteria can exchange genes across "species" through a variety of mechanisms (Sonea & Panisset 1976, Margulis 1981). (2) Reticulate evolution also occurs in plants where allopolyploidy may lead to the instantaneous appearance of a new species possessing the chromosome complement of its two parent species. (3) It is also found in within-species micro-evolution in sexually reproducing eukaryotes. Reticulate patterns may also occur in non-phylogenetic problems such as host-parasite relationships involving host transfer and in the field of ecological biogeography.

2 Algorithm for constructing reticulograms

This section describes the most important features of our reticulogram reconstruction algorithm. A *reticulogram* or *tree network* R is a triplet (E, V, l) where V is a set of vertices, E is a set of edges and l is a *function* of edge lengths assigning real non-negative numbers to the edges. Each vertex i is either an object belonging to a set X or a node belonging to $V - X$. In this study we considered only connected and undirected reticulograms. The algorithm uses as input a dissimilarity matrix $\mathbf{D}$ on the set of n objects and an additive tree T inferred from $\mathbf{D}$ using one of the classical reconstruction algorithms. At each step, the algorithm adds to the additive tree T a new edge (reticulation) of optimal length ensuring the minimisation of the following least-squares loss function:

$$Q = \sum_{i \in X} \sum_{j \in X} (dist(i,j) - d(i,j))^2 \rightarrow min \qquad (1)$$

where $d(i,j)$ is a dissimilarity value between objects i and j, and $dist(i,j)$ is the corresponding value of reticulogram distance defined as a *minimum-path-length distance* between vertices i and j in R.

Makarenkov & Legendre (1999) introduced a statistical criterion Q_1 which measures the gain in fit when a new reticulation is added. The minimum of this criterion provides a stopping rule for addition of reticulations. This function takes into account the least-squares loss function as well as the number of degrees of freedom of the reticulogram under construction:

$$Q_1 = \frac{\sqrt{\sum_{i \in X} \sum_{j \in X} (dist(ij) - d(ij))^2}}{(n(n-1)/2 - N)} = \frac{\sqrt{Q}}{(n(n-1)/2 - N)} \qquad (2)$$

N is the number of edges in the reticulogram. N is equal to $2n-3$ in a binary additive tree with n leaves corresponding to the objects in X and $n-2$ internal nodes. Thus, in this study, the reticulogram will always contain $2n-2$ internal nodes, n of which correspond to the observed objects.

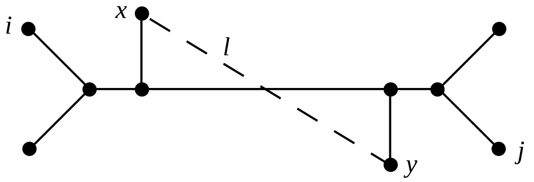

Fig. 1. A new edge of length l can be added to tree T between vertices x and y.

Consider now a binary additive tree T inferred from a dissimilarity d by means of an appropriate fitting method and a pair of vertices x and y in T not linked by an edge (Fig. 1). Using the least-squares loss function, we have to determine an optimal value l for a new edge xy that may be added to the tree T. Let us consider the set $A(xy)$ of all pairs of objects ij of X such that:

$$Min\ \{dist(ix) + dist(jy); dist(jx) + dist(iy)\}\ < dist(ij) \tag{3}$$

The set $A(xy)$ represents the distances between pairs of objects that are susceptible of changing if a new reticulation xy is added. Actually, the set $A(xy)$ can be subdivided into the m subsets $A_1, A_2, ..., A_m$ such that $A(xy) = \{A_1 \cup A_2 \cup ... \cup A_m\}$. They are defined in the following way:
$A_1 = \{ij\}$ such that:
$dist(i,j) - Min\{dist(i,x) + dist(j,y); dist(j,x) + dist(i,y)\} =$
$Min_{\{ij \in A(xy)\}}\{dist(i,j) - Min\{dist(i,x) + dist(j,y); dist(j,x) + dist(i,y)\}\}$
$= l_1$
...
$A_k = \{ij\}$ such that:
$dist(i,j) - Min\{dist(i,x) + dist(j,y); dist(j,x) + dist(i,y)\} = l_k > l_{k-1}$
...
$A_m = \{ij\}$ such that:
$dist(i,j) - Min\{dist(i,x) + dist(j,y); dist(j,x) + dist(i,y)\} =$
$Max_{\{ij \in A(xy)\}}\{dist(i,j) - Min\{dist(i,x) + dist(j,y); dist(j,x) + dist(i,y)\}\}$
$= l_m = dist(x,y) > l_{m-1}$

This subdivision is performed because each different subset A_i can be associated with an interval of possible edge lengths l for which a particular optimisation problem may be formulated. Let us compose a special quadratic function to be minimised for a fixed interval of edge length values. To obtain its optimal solution, suppose that $l_k \leq l \leq l_{k+1}$, where $k = 0...m-1$. This constraint means that if a new edge xy of length l is added to T, only the set of distances $dist(ij)$ such that $ij \in \{A_m \cup A_{m-1} \cup ... \cup A_{k+1}\}$ will change lengths. Thus, the function to minimise to compute the optimal length value of l is as follows:

$$Q^*(xy,k) = \sum_{p=k+1}^{m} \sum_{ij \in A_p} (Min\{dist(i,x) + dist(j,y); dist(j,x) + dist(i,y)\} + l - d(i,j))^2 \rightarrow min \qquad (4)$$

subject to the constraint $l_k \leq l \leq l_{k+1}$. $Q^*(xy,k)$ comprises the quadratic sum of differences between the dissimilarities d and the associated reticulogram distances $dist$, considering only the distances that may change in the reticulogram. The non-trivial solution $l^*(xy,k)$ is (Makarenkov and Legendre, submitted):

$$\frac{\sum_{p=k+1}^{m} \sum_{ij \in A_p} (d(i,j) - Min\{dist(i,x) + dist(j,y); dist(j,x) + dist(i,y)\})}{\sum_{p=k+1}^{m} |A_p|} \qquad (5)$$

This calculation is repeated over all intervals of edge lengths $l_k \leq l \leq l_{k+1}$, for $k = 0...m-1$, for the given pair of vertices xy. The global optimum for criterion Q found for every particular solution, as well as the global optimum of the edge length l over the set of defined intervals, are recursively obtained. To obtain the optimum value for Q over the set of all possible new edges, the computations are repeated for all pairs of tree (reticulogram) vertices not linked by an edge.

3 Application

In a recent study, Makarenkov & Legendre (1999) considered two applications of reticulogram reconstruction. The first one concerned the postglacial dispersal of freshwater fishes in the Québec Peninsula. The second example depicted the morphological differentiation of muskrats in a river valley in Belgium. We will now examine how the method can be applied to represent homoplasy in the phylogenetic tree of primates. Homoplasy is the portion of phylogenetic similarity resulting from convergence. The data, from Hayasaka et al. (1988), consisted of a portion of the protein-coding mitochondrial DNA (898 bases) over 12 species of primates. The dissimilarity matrix (Table 1) was obtained by computing the Hamming distance among the species. First, a phylogenetic tree was inferred from the dissimilarity matrix using the neighbor-joining method (Saitou & Nei 1987). The tree is represented by full lines in Fig. 2. The phylogeny separated four basic groups of primates. The values of criteria Q and Q_1 after approximation of the edge lengths (about this technique, see Makarenkov & Leclerc 1999) were 0.002479 and 0.001106, respectively. Five new edges (reticulations, dashed lines in Fig. 2) were added to the tree by the algorithm. The minimum of Q_1 was reached at the fifth step of the algorithm, which allowed to decrease Q_1 to 0.001041, whereas Q dropped to 0.001733 (gaining about 30%).

	1	2	3	4	5	6	7	8	9	10	11
2 Pan	0.089										
3 Gorilla	0.104	0.106									
4 Pongo	0.161	0.171	0.166								
5 Hylobates	0.182	0.189	0.189	0.188							
6 Macaca fus.	0.232	0.243	0.237	0.244	0.247						
7 M. mulatta	0.233	0.251	0.235	0.247	0.239	0.036					
8 M. fascicul.	0.249	0.268	0.262	0.262	0.257	0.084	0.093				
9 M. sylvan.	0.256	0.249	0.244	0.241	0.242	0.124	0.120	0.123			
10 Saimiri sc.	0.273	0.284	0.271	0.284	0.269	0.289	0.293	0.287	0.287		
11 Tarsius sy.	0.322	0.321	0.314	0.303	0.309	0.314	0.316	0.311	0.319	0.320	
12 Lemur ca.	0.308	0.309	0.293	0.293	0.296	0.282	0.289	0.298	0.287	0.285	0.252

Table 1. Dissimilarity matrix among primates; species 1 is *Homo sapiens*.

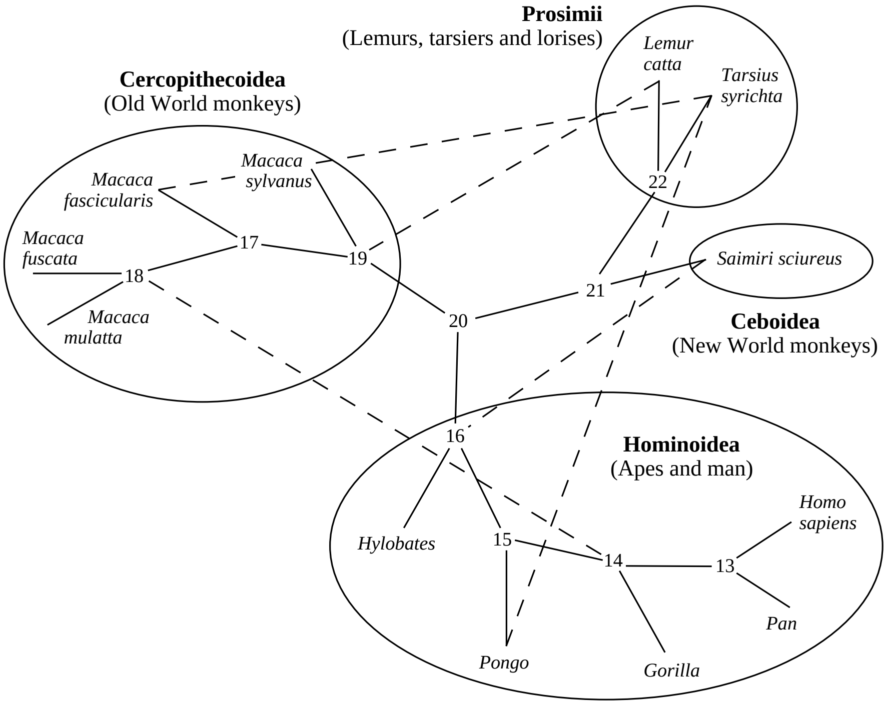

Fig. 2. Reticulogram representing the phylogeny of the primates from Table 1. Full lines: edges of the additive tree. Dashed: reticulations added by algorithm.

How can we interpret reticulations? From the mathematical point of view, each reticulation improves the representation of matrix **D** by the classical additive tree, allowing an optimal gain in fit. From the biological point of view, the lengths of the reticulations are of great importance. If the length of a reticulation is small with respect to the other edge lengths, it may represent a mutation event that occurred during evolution. In the example, the reticulations are long and they occur between distant groups, so that they represent homoplasy (i.e., information representing convergent evolution: parallel evolution and reversals) in the data, which the phylogenetic tree was unable to correctly represent. For instance, the distance between *Homo sapiens* and *Macaca fuscata* is 0.23215 in Table 1. The distance between these species is 0.24133 along the tree, whereas the minimum-path-length reticulogram distance, which includes the reticulation linking the Cercopithecoidea and Hominoidea, is 0.23549. This value is a better representation of the original dissimilarity than the tree path-length distance is.

References

BANDELT, H.-J. (1995): Combination of Data in Phylogenetic Analysis. *Plant Systematics and Evolution Supplementum*, *9*, 355–361.

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

HAYASAKA, K., GOJOBORI, T., and HORAI, S. (1988): Molecular phylogeny and evolution of primate mitochondrial DNA. *Molecular Biology and Evolution*, *5*, 626–644.

KLAUER, K.C. and CARROLL, J.D. (1989): A Mathematical Programming Approach for Fitting General Graphs. *Journal of Classification*, *6*, 247–270.

MAKARENKOV, V. and LECLERC, B. (1999): An Algorithm for the Fitting of a Tree Metric According to a Weighted Least-Squares Criterion. *Journal of Classification*, *16*, 3–27.

MAKARENKOV, V. and LEGENDRE, P. (2000): General Network Representation of a Dissimilarity Matrix: Adding Reticulations to an Additive Tree. *Journal of Classification* (submitted).

MARGULIS, L. (1981): *Symbiosis in Cell Evolution*, San Francisco, CA: W. H. Freeman.

ORTH, B. (1988): Representing Similarities by Distance Graphs: Monotonic Network Analysis (MONA). In: H. H. Bock (Ed.), *Classification and related methods of data analysis*, Amsterdam: North-Holland, 489–494.

SAITOU, N. and NEI, M. (1987): The Neighbor-Joining Method: A New Method for Reconstructing Phylogenetic Trees. *Molecular Biology and Evolution*, *4*, 406–425.

SONEA, S. and PANISSET, M. (1976): Pour une nouvelle bactériologie. *Revue Canadienne de Biologie*, *35*, 103–167.

Makarenkov, V. & P. Legendre. 2000. Improving the additive tree representation of a dissimilarity matrix using reticulations. Pp. 35-40 in: Kiers, H. A. L., J.-P. Rasson, P. J. F. Groenen & M. Schader [eds.] *Data Analysis, Classification, and Related Methods.* Studies in Classification, Data Analysis, and Knowledge Organization. Springer, Berlin. (Proceedings of the IFCS-2000 Conference, Namur, Belgium, 11-14 July 2000.)