Comparison of Four Methods for Inferring Additive Trees from Incomplete Dissimilarity Matrices

Vladimir Makarenkov

Département d'informatique, Université du Québec à Montréal, C.P. 8888, Succ. Centre-Ville, Montréal (Québec), Canada, H3C 3P8.
Institute of Control Sciences, 65 Profsoyuznaya, Moscow 117806, Russia.
(e-mail: makarenkov.vladimir@uqam.ca)

Abstract. The problem of inference of an additive tree from an incomplete dissimilarity matrix is known to be very delicate. As a solution to this problem, it has been suggested either to estimate the missing entries of a given partial dissimilarity matrix prior to tree reconstruction (De Soete, 1984 and Landry et al., 1997) or directly reconstruct an additive tree from incomplete data (Makarenkov and Leclerc, 1999 and Guénoche and Leclerc, 2001). In this paper, I propose a new method, that is based on the least-squares approximation, for inferring additive trees from partial dissimilarity matrices. The capacity of the new method to recover a true tree structure will be compared to those of three well-known techniques for tree reconstruction from partial data. The new method will be proven to work better than widely used Ultrametric and Additive reconstruction techniques, as well as the recently proposed Triangle method on incomplete dissimilarity matrices of different sizes and under different noise conditions.

1 Introduction

Incomplete dissimilarity data can arise in a variety of practical situations. For example, this is often the case in molecular biology, and more precisely in phylogenetics, where an additive or a phylogenetic tree represents an intuitive model of species evolution. The presence of missing data in a distance or dissimilarity matrix among species or taxa can be due to the lack of biological material, imprecision of employed experimental methods, or to a combination of unpredictable factors. Unfortunately, the vast majority of the widely used additive tree fitting techniques, as for example the Neighbor-Joining (Saitou and Nei, 1987), Fitch (Felsenstein, 1997), or BioNJ (Gascuel, 1997) algorithms, cannot be launched unless a complete dissimilarity matrix is available. To solve this challenging problem, some methods have been recently proposed.

There exist in the literature two types of methods, using either *indirect* or *direct* estimation of missing values, for inferring additive trees from incomplete dissimilarity matrices. The first type of methods, or indirect estimation,

relies on the assessing missing cells prior to phylogenetic reconstruction using the properties of path-length matrices representing trees. An additive tree can then be inferred from a complete dissimilarity matrix by means of any available tree-fitting algorithm. The second type of methods handling missing values, or direct estimation, consists of reconstructing a tree directly from an incomplete dissimilarity matrix by using a particular tree-building procedure. As far as the direct estimation techniques are concerned, I have to mention the work by De Soete (1984) and Landry et al. (1996), who showed how to infer additive trees from partial data using either the *ultrametric inequality*:

$$d(i,j) \leq max\{d(i,k); d(j,k)\}, \text{for any } i, j \text{and } k, \tag{1}$$

or the *four-point condition* (Buneman 1971):

$$d(i,j)+d(k,l) \leq max\{d(i,k)+d(j,l); d(i,l)+d(j,k)\}, \text{for any } i, j, k \text{and } l \tag{2}$$

Using the properties of the ultrametric inequality and the four-point condition, one can fill out incomplete matrices; the missing cells can actually be estimated through the combinations of the available ones. As to the direct reconstruction, two tree-building algorithms allowing for missing cells in dissimilarity matrices have been recently proposed by different authors; the Triangle method of Guénoche and Leclerc (2001), see also Guénoche and Grandcolas (1999), relies on a constructive approach, whereas the MW procedure of Makarenkov and Leclerc (1999) is based on a least-squares approximation.

This paper aims first at introducing a new original method for direct reconstruction of additive trees from partial matrices. The second goal consists of proving the efficiency of the proposed method by comparing it to the Ultrametric and Additive indirect procedures, as well as to the Triangle direct reconstruction method. In order to compare the new method to the three above-mentioned existing approaches, Monte Carlo simulations were conducted with dissimilarity matrices of different sizes and with different percentages of missing cells. The performances of the four methods were assessed in terms of both metric and topological recovery. The conducted simulations clearly showed that the new method regularly provided better estimates of the path-length distances between tree leaves, as well as a better recovery of the correct tree topology than the three other competing strategies.

2 Brief description of the new method

The new method for reconstructing trees from partial matrices introduced in this article was inspired by the Method of Weights (MW) proposed in Makarenkov and Leclerc (1999). The latter method used a stepwise addition procedure to infer an additive tree from a complete dissimilarity matrix. The approximation procedure used in the MW was based on a weighted least-squares model. The new method, called *MW-modified*, is an extension of the

MW approach to partial matrices. The first attempt to use the MW method for treatment of partial matrices was made in Levasseur et al.(2000), where the MW procedure was compared to the Triangle method. However, this first attempt to employ the least squares for tree reconstruction from partial matrices showed that the direct MW procedure had to be adjusted to the treatment of missing data.

Let $\mathbf{D}$ be a given dissimilarity matrix on the set X of n taxa. Let us suppose that some entries of $\mathbf{D}$ are missing. The least-squares criterion consists in minimising the following function:

$$Q = \sum_{i<j} (d(i,j) - \delta(i,j))^2, \tag{3}$$

where $\delta(i,j)$ is an obtained estimate for an existing entry $d(i,j)$ of $\mathbf{D}$. The function δ is a tree metric, which is associated with an additive tree; δ verifies the four-point condition.

The following approach was adopted in the MW-modified procedure to build an additive tree from a partial dissimilarity matrix $\mathbf{D}$:

Step 1. The taxa i and j are chosen, such that $d(i,j)$ is a present entry of $\mathbf{D}$. The tree T_2 will comprise the only edge ij of length $d(i,j)$.

Step p, ($\mathbf{p < n}$). Let T_p be an additive tree with p leaves constructed at the previous steps. The leaves of T_p are associated with p taxa from X. Among the $n-p$ taxa from X that are not represented by the leaves of T_p, we have to find one to be placed in the growing tree. We propose to place in T_p, the taxon $p+1$ such that it provides the maximum number of existing dissimilarity entries of type $d(p+1,l)$, where l is a taxon of X already placed in T_p. The exact location of the new leaf $p+1$ in T_{p+1} and the lengths of the three new edges appearing after addition of a new leaf will be determined with respect to the MW procedure (see Makarenkov and Leclerc 1999).

The time complexity of such a new procedure is $O(n^3)$ for a dissimilarity matrix $\mathbf{D}$ of size (nxn). In order to improve the quality of fit in the simulation study described below, I carried out this procedure k times for each dissimilarity matrix, where k was a number of possible existing pairs of taxa i and j to be selected at the first step. Such an exhaustive strategy increases the algorithmic time complexity up to $O(n^5)$, but often enables a substantial improvement in fit.

3 Simulation design

I carried out a series of Monte Carlo simulations to compare the performances of the four competing methods for tree reconstruction from incomplete dissimilarities. Each data set was obtained as follows: first, an unrooted binary tree topology with n leaves and $2n-3$ edges was randomly generated. For each such tree topology, the length of each edge was then selected at random from a uniform distribution on the real interval [0,1], leading to an additive

tree T. The corresponding tree metric tt was computed and normalized to have a unit variance. Four normally distributed random noises with mean zero and, respectively, variances $\sigma^2 = 0.0$ (noise-free condition), 0.1, 0.25, and 0.5, were added to the values of the normalized tree metric tt to obtain variants of the dissimilarity d. Then, 0 to 50 percent of entries were removed from d to obtain a partial dissimilarity used as input for the four tree-reconstructing methods compared in this study. Thus, all considered partial dissimilarity matrices were simulated according to the "tree metric + noise - missing entries" model. For each combination of values $(n, \sigma^2, miss)$, where n (matrix size) = 8, 16, and 24, and mis (percentage of missing values) = 0, 10, 20, 30, 40, and 50, I generated 100 different data sets. However, only the results obtained for $n = 16$ are illustrated in Figures 1 and 2 below.

The metric and a topological recovery provided by each of the four tree-building methods were assessed using the two following quantities:

1. The *proportion of variance accounted for* as expressed in the following formula:

$$\%Var = 100\% \times (1 - \frac{\sum_{i<j}(d(i,j) - \delta(i,j))^2}{\sum_{i<j}(d(i,j) - m(d))^2}), \tag{4}$$

where $m(d)$ is the mean value of the initial dissimilarity d, and δ is the obtained tree metric.

2. The *Robinson and Foulds topological distance* (Robinson and Foulds 1981) between the true tree T and the solution tree corresponding to the obtained tree metric d was also considered. This important criterion of tree similarity is equal to the minimum number of elementary operations, consisting of merging or splitting vertices, necessary to transform one tree into the other.

In this Monte Carlo study, I used the additive tree generating strategy that was also employed in Makarenkov and Leclerc (1999), and Makarenkov and Legendre (2001). Another way of simulating data for the additive model was suggested by Lausen and Degens (1988).

4 Conclusion

In this paper, the performances of four different tree-building methods for incomplete dissimilarity matrices were compared. I carried out a Monte Carlo study to determine which method provides better metric and topological recoveries under different noise conditions and for different percentages of missing values. When analyzing curves illustrated in Figures 1 and 2, one can notice that the proposed MW-modified procedure generally achieves better results in terms of both metric and topological recovery than the three other methods. The new method is particularly good in case of complete absence of noise (noise = 0, in Figures 1 and 2). A number of interesting trends can be observed when examining curves behaviour in Figure 1 and 2. As to the percentage of variance: the Additive procedure becomes very unstable when

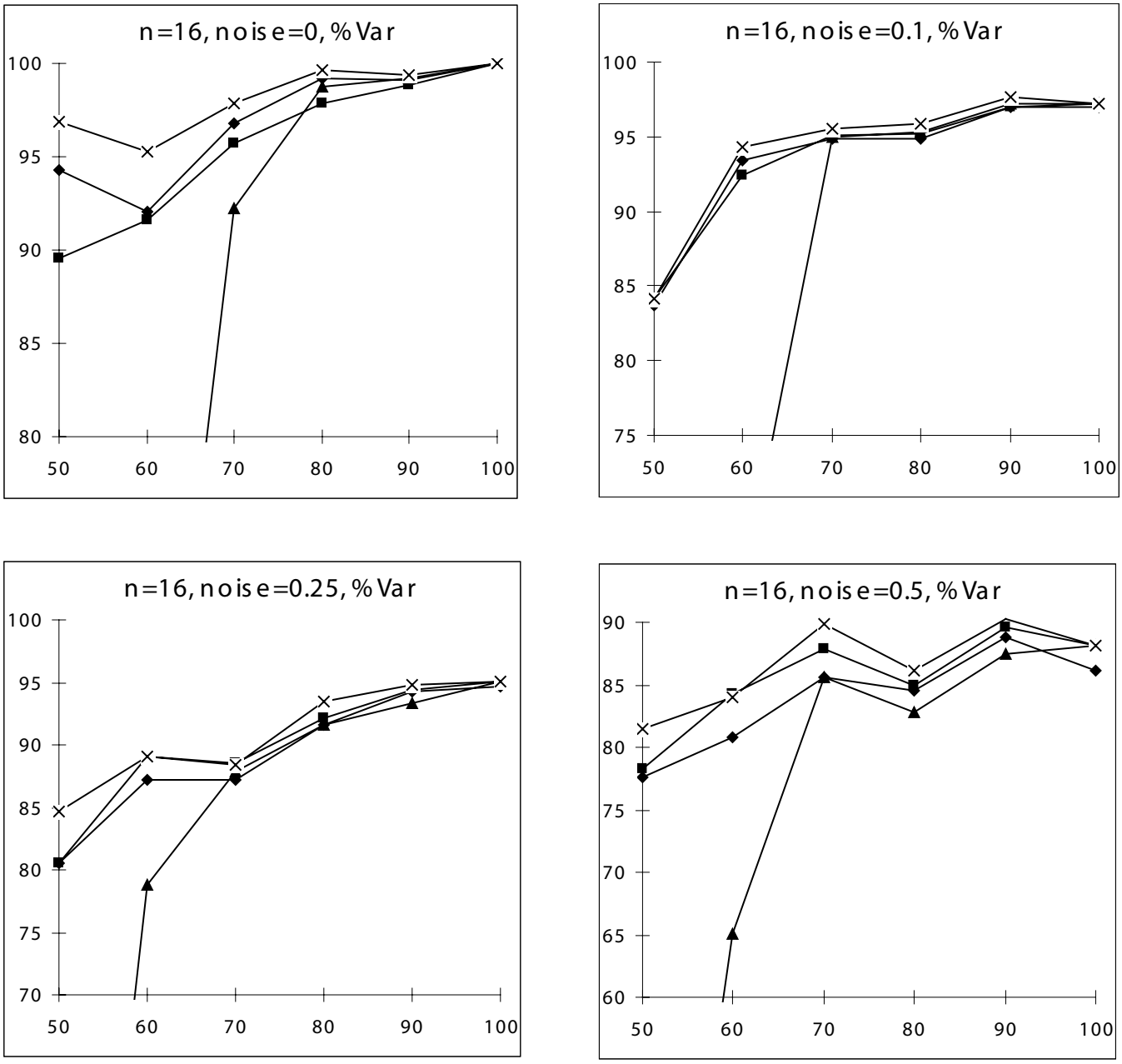

Fig. 1: Metric recovery values obtained for different percentages of missing entries and under four different noise conditions. The four competing methods [Triangle ♦, Ultrametric ■, Additive ▲, and MW-modified X] were tested on dissimilarity matrices of size (16x16). The abscissa axis represents the percentage of existing entries in a given dissimilarity matrix; the ordinate axis represents the percentage of variance of a given dissimilarity accounted for by an obtained tree metric. For each case, the mean values (over 100 simulated data sets) of the percentage of variance are given. Larger values of the percentage of variance accounted for point out a better recovery achieved by a tree reconstruction method.

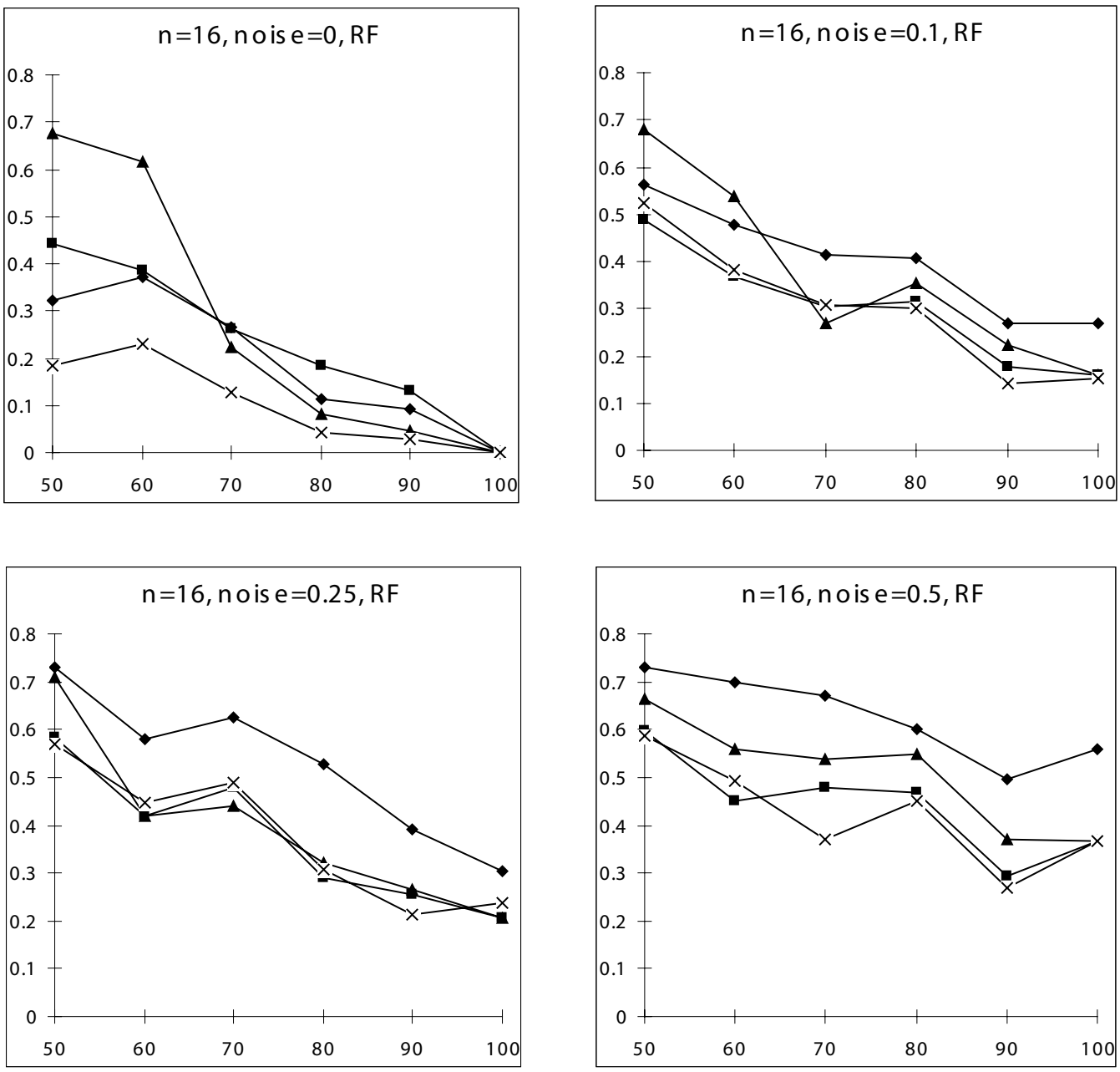

Fig. 2: Topological recovery values obtained for different percentages of missing entries and under four different noise conditions. The four competing methods [Triangle ♦, Ultrametric ■, Additive ▲, and MW-modified X] were tested on dissimilarity matrices of size (16x16). The abscissa axis represents the percentage of existing entries in a given dissimilarity matrix; the ordinate axis represents the Robinson and Foulds (RF) topological distance between a given true additive tree and an obtained tree. The computed RF distance was normalized by its largest value, which is 2n-6 for two binary additive trees with n leaves. For each case, the mean values (over 100 data sets) of the topological distance are given. Lower values of the RF distance point out a better recovery achieved by a tree reconstruction

more than 30 percent of entries are missing in a given dissimilarity matrix; the results provided by the Triangle method and the Ultrametric procedure are very close; the MW-modified procedure slightly outperforms both latter methods in the most of the situations. As to the normalized Robinson and Foulds topological distance: the Triangle method does not resist well to the increasing of noise; the best results are regularly obtained by the MW-modified method or the Ultrametric procedure; similarly to the percentage of variance, the Additive procedure can not cope well with big percentage of missing cells in a given dissimilarity matrix. Surprisingly, the Ultrametric procedure which does not provide good results for noise-free data, the same conclusion was also made by Landry at al. (1996) and Guénoche and Grandcolas (1999), becomes much more competitive when the noise increases. Similar trends were observed for additive trees with 8 and 24 leaves, for which the detailed results were not presented in this paper. Ultimately, I would recommend using the MW-modified method for treatment of data that are free of noise and the MW-modified method or the Ultrametric procedure for processing "noisy" data.

5 Software

The four methods compared in the framework of the present study were implemented in T-Rex (*tree and reticulogram reconstruction*) package intended for reconstructing additive trees and reticulation networks (Makarenkov, 2001). This computer application also includes a number of well-known tree fitting methods, as well as some methods for modelling reticulation networks between considered species or taxa. A tree structure obtained by means of one of these methods can be visualized using Hierarchical, Radial, or Axial drawing and then manipulated interactively. The Windows and Macintosh versions of this software were made freely available for researchers at the T-Rex web site at http://www.fas.umontreal.ca/biol/casgrain/en/labo/t-rex.

6 Acknowledgement

The author is grateful to Alain Guénoche and François-Joseph Lapointe for providing me with the source code of the Triangle, Ultrametric, and Additive methods.

References

BUNEMAN, P. (1971): The Recovery of Trees From Measures of Dissimilarity. In: F.R. Hudson, D.G. Kendall and P. Tautu (Eds.): *Mathematics in Archeological and Historical Sciences*. Edinburgh University Press, Edinburgh, 387–395.

DE SOETE, G. (1983): Additive-tree representations of incomplete dissimilarity data. *Quality and Quantity, 18*, 387–393.

FELSENSTEIN, J. (1997): An alternating least-squares approach to inferring phylogenies from pairwise distances. *Systematic Zoology, 46*, 101–111.

GASCUEL, O. (1997): BIONJ: an improved version of the NJ algorithm based on a simple model of sequence data. *Molecular Biology and Evolution, 14*, 685–695.

GUÉNOCHE, A. and LECLERC, B. (2001): The triangle method to build X-trees from incomplete distance matrices. *RAIRO Operations Research, 35*, 283–300.

GUÉNOCHE, A. and GRANDCOLAS, S. (1999): Approximation par arbre d'une distance partielle. *Mathématiques,Informatique et Sciences Humaines, 146*, 51–64.

LANDRY, P. A., LAPOINTE, F.-J., and KIRSCH, J. A. W. (1996): Estimating phylogenies from distance matrices: additive is superior to ultrametric estimation. *Molecular Biology and Evolution, 13*, 818–823.

LAUSEN, B. and DEGENS, P. O. (1988): Evaluation of the reconstruction of phylogenies with DNA-DNA hybridization data. In Bock, H. H. (ed.), *Classification and related methods of data analysis*, North Holland, Amsterdam, 367–374.

LEVASSEUR, C., LANDRY, P. A., and LAPOINTE, F.-J. (2000): Estimating trees from incomplete distance matrices: a comparison of two methods. In Kiers, H. A. L., Rasson, J.-P., Groenen, P. J. F. and Schader, M. (eds.), *Data analysis, Classification and Related Methods*, Springer, 149–154.

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*, 3–26.

MAKARENKOV, V. and LEGENDRE, P. (2001): Optimal Variable Weighting for Ultrametric and Additive Trees and K-means Partitioning: Methods and Software. *Journal of Classification, 18*, 245–271.

MAKARENKOV, V. (2001): T-Rex: reconstructing and visualizing phylogenetic trees and reticulation networks, *Bioinformatics, 17*, 664–668.

ROBINSON, D.R. and FOULDS, L.R. (1981): Comparison of phylogenetic trees. *Mathematical Biosciences, 53*, 131–147.

SAITOU, N. and NEI, M. (1987): The neighbor-joining method: a new method for reconstructing phylogenetic trees. *Molecular Biology and Evolution, 4*, 406–425.